April 1991 LIDS-R-2030

LARGE DEVIATIONS AND APPLICATIONS: THE FINITE DIMENSIONAL CASE

A. Dembo
Department of Mathematics and Department of Statistics
Stanford University
Stanford, CA 94305, USA

O. Zeitouni
Department of Electrical Engineering, Technion
Haifa 32000, Israel

This research was supported by the Air Force Office of Scientific Research under contract AFOSR-89-0276B and by the Army Research Office under contract DAAL03-86-K-0171.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 1991		**00-04-1991 to 00-04-1991**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Large Deviations and Applications: The Finite Dimensional Case	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**99**	

Preface and Warning

These notes, which form the first chapter of a forthcoming book, are intended to serve as lecture notes on the topic of large deviations and applications for students whose background and interests are in applications which involve finite dimensional spaces. Although narrow in their scope, these notes present a good deal of the methods available for more general situations. A glaring omission is the method of sub-additivity, which will be discussed in another chapter in the book. Another deficiency of these notes is the sketchy bibliography and historical. We hope to correct this in the book.

The funny looking ?? appearing in various places mean references to later chapters. Please disregard those.

Acknowledgments Special thanks go to Dan Stroock who taught one of us (O.Z.) large deviations theory and provided many comments and useful insights. This course was taught at Stanford and the Technion. The comments of people who sat in these courses, and in particular the comments of Yuval Peres, contributed much to correct mistakes and omissions. We wish to thank Sam Karlin for suggesting the application of rare long segments in random walks (section 2.5). One of us (A.D.) thanks Tom Cover and Joy Thomas for a preprint of their forthcoming book which influenced our treatment of sections 2.1.1 and 2.7.

Contents

Chapter 1

Introduction

1.1 The Large Deviations Principle

A *large deviations principle* characterizes the limiting behavior (as $\epsilon \to 0$) of a family of probability measures μ_ϵ on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ in terms of a *rate function.* This characterization is via 'tight' asymptotic upper and lower exponential bounds on the values that μ_ϵ assigns to close and open subsets of $\mathcal{X}$ (respectively). For that purpose $\mathcal{X}$ should be a topological space so that open and closed subsets of $\mathcal{X}$ are well defined concepts and as usual only measurable sets (i.e., elements of $\mathcal{B}_\mathcal{X}$, the Borel sigma field on $\mathcal{X}$) are of interest.

Definitions: A **rate function** I is any mapping $I : \mathcal{X} \to [0, \infty]$ such that for any $\alpha \in [0, \infty)$ the level set $\Psi_I(\alpha) \triangleq \{x : I(x) \leq \alpha\}$ is a closed subset of $\mathcal{X}$. A **good rate function** is a rate function for which all the level sets $\Psi_I(\alpha)$ are compact subsets of $\mathcal{X}$.

Alternatively, a rate function is any non-negative, lower semicontinuous function on $\mathcal{X}$ (for the definitions and a proof of this fact see Appendix ??). Throughout, let $\mathcal{D}_I$ denote the set of points in $\mathcal{X}$ of *finite rate*, namely $\mathcal{D}_I \triangleq \{x : I(x) < \infty\}$.

Definition: We say that μ_ϵ **satisfies the large deviations principle with rate function** I if, for all $\Gamma \in \mathcal{B}_\mathcal{X}$,

$$-\inf_{x \in \Gamma^o} I(x) \leq \liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(\Gamma) \leq \limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x) \tag{1.1.1}$$

where $\overline{\Gamma}$ is the closure of Γ, Γ^o is the interior of Γ and the infimum of I over an empty set is

interpreted throughout as ∞.

Corollary 1.1.1 *When μ_ϵ satisfies a large deviations principle with rate function I and $\Gamma \in \mathcal{B}_\mathcal{X}$ is such that*

$$\inf_{x \in \Gamma^\circ} I(x) = \inf_{x \in \overline{\Gamma}} I(x) \triangleq I_\Gamma \tag{1.1.2}$$

then

$$\lim_{\epsilon \to 0} \epsilon \log \mu_\epsilon(\Gamma) = -I_\Gamma \tag{1.1.3}$$

Remark: When I is a good rate function and $\overline{\Gamma} \cap \mathcal{D}_I$ is non-empty then there exists at least one point $x^* \in \overline{\Gamma}$ where $\inf_{x \in \overline{\Gamma}} I(x)$ is achieved.

A set Γ which satisfies (1.1.2) is called an I **Continuity Set.** In general, a large deviations principle implies a precise limit as in (1.1.3) only for I continuity sets.

In particular, points are typically not open subsets of $\mathcal{X}$ so a large deviations principle *does not* result with an asymptotic exponential estimate on the probability that μ_ϵ assigns to each point in $\mathcal{X}$. Better results may well be derived on a case by case basis for specific families of measures μ_ϵ and particular sets. While such results do not fall within our definition of a large deviations principle few illustrative examples are included in this book (see Sections 2.1,2.6,2.10).

An alternative formulation of the large deviations principle is as follows:

(a). For any closed set $F \in \mathcal{B}_\mathcal{X}$,

$$\limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(F) \leq - \inf_{x \in F} I(x) \tag{1.1.4}$$

(b). For any open set $G \in \mathcal{B}_\mathcal{X}$,

$$\liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(G) \geq - \inf_{x \in G} I(x) \tag{1.1.5}$$

The inequality (1.1.4) is also called the large deviations upper bound while (1.1.5) is called the large deviations lower bound. Thus, the large deviations principle corresponds to the scenario in which both bounds hold with the same rate function.

A few observations are now in place. First, observe that since $\mu_\epsilon(\mathcal{X}) = 1$ for any ϵ it is necessary that $\inf_{x \in \mathcal{X}} I(x) = 0$ for the upper bound to hold (and when I is a good rate function then there

exists at least one point x in which $I(x) = 0$). Next, observe that the upper bound trivially holds whenever $\inf_{x \in F} I(x) = 0$ while the lower bound trivially holds whenever $\inf_{x \in G} I(x) = \infty$. This leads to another alternative formulation of the large deviations principle which is useful when trying to prove such a principle.

(a). For any closed set $F \in \mathcal{B}_{\mathcal{X}}$ contained in the complement of $\Psi_I(\alpha)$,

$$\limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(F) \leq -\alpha \tag{1.1.6}$$

(b). For any $x \in \mathcal{D}_I$ and any open neighborhood $G \in \mathcal{B}_{\mathcal{X}}$ of x in $\mathcal{X}$,

$$\liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(G) \geq -I(x) \tag{1.1.7}$$

The inequality (1.1.7) reveals the local nature of the lower bound which should only be proved for "small open sets". Moreover, the following indicates that the upper bound may be proved first for an approximate I functional:

Definition: For any $\delta > 0$ let $I^\delta(x) = I(x) - \delta$ when $x \in \mathcal{D}_I$ and $I^\delta(x) \stackrel{\Delta}{=} \frac{1}{\delta}$ when $x \in \mathcal{D}_I^c$.

Since for any set F

$$\lim_{\delta \to 0} \inf_{x \in F} I^\delta(x) = \inf_{x \in F} I(x)\,, \tag{1.1.8}$$

it suffices to prove that for any $\delta > 0$ and for any closed set F

$$\limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(F) \leq -\inf_{x \in F} I^\delta(x) \tag{1.1.9}$$

in order to conclude that the upper bound (1.1.4) holds.

Actually, a common technique for proving the **existence** of a large deviations principle is by implicitly defining

$$I(x) \stackrel{\Delta}{=} \sup_{\{A \in \mathcal{C}^o : x \in A\}} \mathcal{L}_A \tag{1.1.10}$$

where

$$\mathcal{L}_A \stackrel{\Delta}{=} -\liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A)\,, \tag{1.1.11}$$

and $\mathcal{C}^o \subset \mathcal{B}_{\mathcal{X}}$ is any basis for the topology of $\mathcal{X}$ (namely, any open set G is the union of sets from $\mathcal{C}^o$).

This definition results with a rate function I for which the lower bound (1.1.5) holds. Moreover, when $\lim_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A)$ *exists* for any $A \in \mathcal{C}^o$, then the upper bound (1.1.4) also holds for any compact $F \in \mathcal{B}_\mathcal{X}$ (these statements are proved in Section ??).

This approach in which the upper bound is first proved directly for compact sets and then its validity extended to all closed sets serves as a motivation for the following definition.

Definition: A sequence of probability measures μ_ϵ satisfies **weakly** the large deviations principle with rate function I if the upper bound (1.1.4) holds for F **compact** and the lower bound (1.1.5) holds.

The following auxiliary *exponential tightness* property suffices for extending the ***weak*** large deviations principle to a full large deviations principle with a good rate function (this is shown in the sequel).

Definition: A family of measures μ_ϵ is **exponentially tight** if for any $\alpha < \infty$ there exists a compact set $K_\alpha \in \mathcal{B}_\mathcal{X}$ such that

$$\limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(K_\alpha^c) < -\alpha \tag{1.1.12}$$

where K_α^c denotes the complement of the set K_α.

Remark: The measures μ_ϵ may satisfy a large deviations principle with a good rate function without being exponentially tight. Beware of this common logical mistake.

The exponential tightness (and the alternative statement of the upper bound (1.1.9) are applied in the following lemma for strengthening a weak large deviations principle.

Lemma 1.1.1 *Let μ_ϵ be an exponentially tight family of probability measures.*
(a). If the upper bound (1.1.4) holds for all compact sets then it also holds for all closed sets.
(b). If the lower bound (1.1.5) holds for all open sets then $I(\cdot)$ is a good rate function.

Thus, when an exponentially tight family of measures satisfies weakly the large deviations principle with a rate function $I(\cdot)$ then I is a good rate function and a full large deviations principle holds.
Proof: (a). Consider an arbitrary closed set F. All we need is to establish (1.1.6) whenever $F \subset \Psi_I(\alpha)^c$. Fix any such $\alpha < \infty$. Clearly,

$$\mu_\epsilon(F) \leq \mu_\epsilon(F \cap K_\alpha) + \mu_\epsilon(K_\alpha^c) \,,$$

where K_α is the compact set in (1.1.12). Part (a) of the lemma now follows by the inequality (1.1.12) and the upper bound (1.1.4) for the compact set $F \cap K_\alpha$ (note that $F \cap K_\alpha \subset \Psi_I(\alpha)^c$, so $\inf_{x \in F \cap K_\alpha} I(x) \geq \alpha$).

(b). Applying the lower bound (1.1.5) to the open set K_α^c, one concludes by (1.1.12) that $\inf_{x \in K_\alpha^c} I(x) > \alpha$. Therefore, $\Psi_I(\alpha) \subseteq K_\alpha$, yielding the compactness of the closed level set $\Psi_I(\alpha)$. As this argument holds for any $\alpha < \infty$, it follows that $I(\cdot)$ is a good rate function. □

A countable family of measures μ_n is considered in many cases (for example when μ_n is the law governing the empirical mean of n random variables). Then, a large deviations principle corresponds to the statement

$$-\inf_{x \in \Gamma^o} I(x) \leq \liminf_{n \to \infty} a_n \log \mu_n(\Gamma) \leq \limsup_{n \to \infty} a_n \log \mu_n(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x) \,, \tag{1.1.13}$$

for some sequence $a_n \to 0$. Note that here a_n replaces ϵ of (1.1.1) and similarly the statements (1.1.4)-(1.1.7) may be appropriately modified.

1.2 Major large deviations techniques

The major disadvantage of the very general approach outlined above is that it does not really reveal what the rate function values are. However, since the rate function associated with a large deviations principle is, under mild conditions, unique (as proved in Section ??) the following two step method is quite useful:

(a). Prove the existence of a large deviations principle.

(b). Verify certain properties of the rate function (typically convexity and/or smoothness) and from them deduce a more convenient (explicit) characterization of this function.

Indeed this method prevails to a certain degree throughout Chapters ??-?? of this book.

Nevertheless, when $\mathcal{X}$ is a subset of a finite dimensional vector space (specifically when $\mathcal{X} \subset \mathbb{R}^d$ for some $d < \infty$) then one can typically prove directly the existence of a large deviations principle with an **explicit** rate function. Chapter 2 of this book which is dedicated to this class of problems serves at least three purposes:

(a). To illustrate the sharpest possible results and emphasize both the reason for the existence of a large deviations principle and the types of rate functions one expects to find.

(b). To demonstrate the different methods for proving explicit large deviations statements in simple scenarios when one's eyesight is not yet obscured by various technical details.

(c). To present an important class of interesting results and their applications while requiring relatively little mathematical background.

Chapter 2

Large Deviations principles for finite dimensional spaces

2.1 Combinatorial Techniques for finite alphabet

Throughout this section all random variables assume values in the finite set $\Sigma = \{a_1, a_2, \ldots, a_{|\Sigma|}\}$ (Σ is also sometimes called the underlying alphabet). Combinatorial methods are then applicable for deriving large deviations principles for empirical measures (see Sections 2.1.1 and 2.1.3 below) and for empirical means (see Section 2.1.2 below). While the scope of these methods is limited to finite alphabets, they illustrate the results one can hope to (and should indeed) derive for more abstract alphabets. Some of these results are actually direct consequences of the large deviations principles derived below via the combinatorial method. For example, in Sections ?? and ?? Theorems 2.1.1 and 2.1.3 are proved for a rather abstract alphabet Σ (specifically, for any Hausdorff topological space Σ). The combinatorial methods, unlike all other approaches for deriving large deviations principles, are based upon point estimates. For example they bound the probabilities associated with each possible outcome of the empirical measure (also denoted as *type*, see the definition below). This turns out to be very useful for some statistical applications (an example is presented in Section 2.6).

2.1.1 The method of types and Sanov's theorem

Let $X_1, X_2 \ldots, X_n$ be a sequence of random variables (*r.v.*) which are independent, identically distributed (*i.i.d.*) according to the law $\mu \in M_1(\Sigma)$. Throughout, $M_1(\Sigma)$ denotes the space of all probability measures (laws) on the alphabet Σ. Typically we shall specify the topology with which $M_1(\Sigma)$ is equipped. However, recall that here Σ is a finite set, so $M_1(\Sigma)$ is identified with the standard probability simplex in $\mathbb{R}^{|\Sigma|}$, i.e, the set of all real valued vectors with $|\Sigma|$ non-negative components which sum to 1, and open sets in $M_1(\Sigma)$ are induced by open sets in $\mathbb{R}^{|\Sigma|}$.

Let Σ_μ denote the support of the law μ, i.e., the set $\Sigma_\mu \stackrel{\Delta}{=} \{a_i : \mu(a_i) > 0\}$. In general, Σ_μ may be a strict subset of Σ (i.e., μ may assign zero probability to some of the symbols in Σ). However, when considering one underlying measure μ we may without loss of generality (w.l.o.g.) reduce Σ to Σ_μ by ignoring all symbols which appear with zero probability. This indeed is implicitly assumed throughout this Section (in Section 2.6 below we encounter a scenario where we have to keep track of various support sets of the form of Σ_μ).

Definition 2.1.1 *The type $L_n^{\mathbf{x}}$ of a sequence $\mathbf{x} \stackrel{\Delta}{=} \{x_1, \ldots, x_n\}$ is the empirical measure (law) associated with this sequence. Specifically, $L_n^{\mathbf{x}} = \{L_n^{\mathbf{x}}(a_1), \ldots, L_n^{\mathbf{x}}(a_{|\Sigma|})\}$ is an element of $M_1(\Sigma)$, where*

$$L_n^{\mathbf{x}}(a_i) = \frac{1}{n} \sum_{j=1}^{n} 1_{x_j = a_i} \stackrel{\Delta}{=} \frac{1}{n} N(a_i|\mathbf{x}) \,, \quad i = 1, \ldots, |\Sigma| \tag{2.1.1}$$

Let $\mathcal{L}_n$ denote the set of all possible types of sequences of length n. Thus, $\mathcal{L}_n \stackrel{\Delta}{=} \{\nu : \nu = L_n^{\mathbf{x}} \text{ for some } \mathbf{x}\} \subset \mathbb{R}^{|\Sigma|}$. Note that the *random* empirical measure $L_n^{\mathbf{X}}$ associated with the sequence $\mathbf{X} \stackrel{\Delta}{=} \{X_1, \ldots, X_n\}$ must be an element of the set $\mathcal{L}_n$.

The usefulness of the notion of types for finite alphabets as well as the reason why it is not readily extended to more general alphabets are due to the following 'volume' and 'approximation distance' estimates:

Lemma 2.1.1 *(a).* $|\mathcal{L}_n| \leq (n+1)^{|\Sigma|}$

(b). For any probability vector $\nu \in M_1(\Sigma)$

$$d_V(\nu, \mathcal{L}_n) \stackrel{\Delta}{=} \inf_{\nu' \in \mathcal{L}_n} d_V(\nu, \nu') \leq \frac{|\Sigma|}{2n} \,, \tag{2.1.2}$$

where $d_V(\nu,\nu') \stackrel{\Delta}{=} \sup_{A\subset\Sigma}[\nu(A)-\nu'(A)]$ *is the variational distance between the measures* ν *and* ν'.

Proof: Note that any component of the vector $L_n^{\mathbf{x}}$ belongs to the set $\{\frac{0}{n},\frac{1}{n},\ldots,\frac{n}{n}\}$ whose cardinality (size) is $(n+1)$. Part (a) of the lemma follows since the vector $L_n^{\mathbf{x}}$ is specified by $|\Sigma|$ such quantities each of which assumes at most $(n+1)$ distinct values.

To prove part (b) observe that $\mathcal{L}_n$ indeed contains all probability vectors which are composed of $|\Sigma|$ elements from the set $\{\frac{0}{n},\frac{1}{n},\ldots,\frac{n}{n}\}$. Thus, for any vector $\nu\in M_1(\Sigma)$ there exists a vector $\nu'\in\mathcal{L}_n$ with $|\nu(a_i)-\nu'(a_i)|\leq\frac{1}{n}$ for $i=1,\ldots,|\Sigma|$. The bound of (2.1.2) now follows since for any discrete alphabet

$$d_V(\nu,\nu')\leq\frac{1}{2}\sum_i|\nu(a_i)-\nu'(a_i)|\,.$$

□

Remarks:

1. Actually, since $L_n^{\mathbf{x}}$ is a probability vector, at most $|\Sigma|-1$ such quantities should be specified and so $|\mathcal{L}_n|\leq(n+1)^{|\Sigma|-1}$.

2. Lemma 2.1.1 states that the volume of $\mathcal{L}_n$, the support of the random empirical measures $L_n^{\mathbf{X}}$, grows polynomially in n and further that for large enough n, the set $\mathcal{L}_n$ approximates uniformly and arbitrarily well (in the sense of variational distance) any measure in $M_1(\Sigma)$. Both properties are invalid when $|\Sigma|=\infty$.

Definition 2.1.2 *The type (composition) class* $T(\nu)$ *of a probability law* $\nu\in\mathcal{L}_n$ *is the set* $T(\nu)=\{\mathbf{x}\in\Sigma^n\,:\,L_n^{\mathbf{x}}=\nu\}$.

Outcomes $\mathbf{x}$ in the same type class are equally likely when $X_1,\ldots,X_n$ are *i.i.d.* random variables. This is part of the first among the three lemmas in the sequel which yield a strong prelude to large deviations results by estimating the exponential growth of each type class and the precise probability of any specific outcome from a given type class. The following definitions are useful for that purpose:

Definitions:

(1). The entropy of a probability vector ν is $H(\nu) \stackrel{\Delta}{=} -\sum_{i=1}^{|\Sigma|} \nu(a_i) \log \nu(a_i)$, where $0 \log 0 = 0$ throughout.

(2). The (relative) entropy of a probability vector ν relative to μ is

$$H(\nu|\mu) \stackrel{\Delta}{=} \sum_{a_i \in \Sigma_\nu} \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} ,$$

when $\Sigma_\nu \subseteq \Sigma_\mu$ (in which case $H(\nu|\mu) < \infty$). Otherwise $H(\nu|\mu) \stackrel{\Delta}{=} \infty$ (whenever $\nu(a_i) > 0$ while $\mu(a_i) = 0$ for some $a_i \in \Sigma$).

Remark: Note that $H(\cdot|\mu)$ is a rate function. Indeed, the set $M_1(\Sigma_\mu)$ is a closed subset of $M_1(\Sigma)$ in which $H(\cdot|\mu)$ is a continuous function (note that the function $-x \log x$ is continuous on [0,1]), while $H(\cdot|\mu) = \infty$ outside $M_1(\Sigma_\mu)$. Finally, the non-negativity of $H(\cdot|\mu)$ follows by Jensen's inequality.

Lemma 2.1.2 *If* $\mathbf{x} \in T(\nu)$ *for* $\nu \in \mathcal{L}_n$, *then*

$$\text{Prob}_\mu(\{X_1, \ldots . X_n\} = \mathbf{x}) = e^{-n[H(\nu)+H(\nu|\mu)]} \tag{2.1.3}$$

where Prob_μ *corresponds to the probability law* μ^{Z_+} *associated with an infinite sequence of random variables* $\{X_j\}$ *which are independent and identically distributed according to* $\mu \in M_1(\Sigma)$.

Proof: Clearly (2.1.3) holds when $H(\nu|\mu) = \infty$ as the random empirical measure $L_n^{\mathbf{X}}$ concentrates on types $\nu \in \mathcal{L}_n$ for which $\Sigma_\nu \subseteq \Sigma_\mu$ (i.e, $H(\nu|\mu) < \infty$). We may thus assume that $H(\nu|\mu) < \infty$ and $\Sigma_\nu \subseteq \Sigma_\mu$.

Since $L_n^{\mathbf{x}} = \nu$ and $H(\nu) + H(\nu|\mu) = -\sum_{i=1}^{|\Sigma|} \nu(a_i) \log \mu(a_i)$ it follows that

$$\text{Prob}_\mu(\{X_1, \ldots, X_n\} = \mathbf{x}) = \prod_{i=1}^{|\Sigma|} \mu(a_i)^{N(a_i|\mathbf{x})} = \prod_{i=1}^{|\Sigma|} \mu(a_i)^{n\nu(a_i)} = e^{-n[H(\nu)+H(\nu|\mu)]} \tag{2.1.4}$$

□

In particular, since $H(\mu|\mu) = 0$, when $\mu \in \mathcal{L}_n$ and $\mathbf{x} \in T(\mu)$ then

$$\text{Prob}_\mu(\{X_1, \ldots, X_n\} = \mathbf{x}) = e^{-nH(\mu)} . \tag{2.1.5}$$

Lemma 2.1.3 *For any* $\nu \in \mathcal{L}_n$, $\quad (n+1)^{-|\Sigma|} e^{nH(\nu)} \leq |T(\nu)| \leq e^{nH(\nu)}$

Remark: Since $|T(\nu)| = \left(\begin{array}{c} n \\ n\nu(a_1), \ldots, n\nu(a_{|\Sigma|}) \end{array} \right)$, one might get a good estimate of $|T(\nu)|$ by applying Stirling's approximation (see [12], pg 48). We take here a somewhat different route, of a more information theoretic flavor.

Proof: Any type class has probability at most one and all its members are of equal probability. Therefore, for any $\nu \in \mathcal{L}_n$, by (2.1.5),

$$1 \geq \text{Prob}_\nu(L_n^{\mathbf{X}} = \nu) = \text{Prob}_\nu(\mathbf{X} \in T(\nu)) = e^{-nH(\nu)}|T(\nu)| \tag{2.1.6}$$

and the upper bound on $|T(\nu)|$ follows. The lower bound follows from the inequality (which is proved below)

$$\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu) \geq \text{Prob}_\nu(L_n^{\mathbf{X}} = \nu'), \quad \forall \nu', \nu \in \mathcal{L}_n, \tag{2.1.7}$$

as thus

$$1 = \sum_{\nu' \in \mathcal{L}_n} \text{Prob}_\nu(L_n^{\mathbf{X}} = \nu') \leq |\mathcal{L}_n| \, \text{Prob}_\nu(L_n^{\mathbf{X}} = \nu) = |\mathcal{L}_n| e^{-nH(\nu)} \, |T(\nu)| \, ,$$

while $|\mathcal{L}_n| \leq (n+1)^{|\Sigma|}$ by Lemma 2.1.1.

It suffices to prove the inequality (2.1.7) for $\nu' \in \mathcal{L}_n$ such that $\Sigma_{\nu'} \subset \Sigma_\nu$ (as otherwise $\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu') = 0$). Thus, without loss of generality one may assume that $\Sigma = \Sigma_\nu$. Now, consider the ratio

$$\frac{\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu)}{\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu')} = \frac{\left(\begin{array}{c} n \\ n\nu(a_1) \ldots n\nu(a_{|\Sigma|}) \end{array} \right)}{\left(\begin{array}{c} n \\ n\nu'(a_1) \ldots n\nu'(a_{|\Sigma|}) \end{array} \right)} \frac{\prod_{i=1}^{|\Sigma|} \nu(a_i)^{n\nu(a_i)}}{\prod_{i=1}^{|\Sigma|} \nu(a_i)^{n\nu'(a_i)}} =$$

$$= \prod_{i=1}^{|\Sigma|} \frac{(n\nu'(a_i))!}{(n\nu(a_i))!} \nu(a_i)^{n\nu(a_i) - n\nu'(a_i)} \, . \tag{2.1.8}$$

This last expression is a product of terms of the form $\frac{m!}{l!}\left(\frac{l}{n}\right)^{l-m}$. By induction $\frac{m!}{l!} \geq l^{(m-l)}$ for any $m, l \in Z^+$, and thus (2.1.8) yields

$$\frac{\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu)}{\text{Prob}_\nu(L_n^{\mathbf{X}} = \nu')} \geq \prod_{i=1}^{|\Sigma|} n^{n\nu'(a_i) - n\nu(a_i)} = n^{n\left[\sum_{i=1}^{|\Sigma|} \nu'(a_i) - \sum_{i=1}^{|\Sigma|} \nu(a_i)\right]} = 1 \tag{2.1.9}$$

□

Lemma 2.1.4 (Large deviations probabilities) *For any $\nu \in \mathcal{L}_n$*

$$(n+1)^{-|\Sigma|}\, e^{-nH(\nu|\mu)} \leq \mathrm{Prob}_\mu(L_n^{\mathbf{X}} = \nu) \leq e^{-nH(\nu|\mu)} \tag{2.1.10}$$

Proof: By Lemma 2.1.2

$$\begin{aligned} \mathrm{Prob}_\mu(L_n^{\mathbf{X}} = \nu) &= |T(\nu)|\, \mathrm{Prob}_\mu(\{X_1, \ldots, X_n\} = \mathbf{x} > \; L_n^{\mathbf{x}} = \nu) \\ &= |T(\nu)|\, e^{-n[H(\nu)+H(\nu|\mu)]} \; . \end{aligned} \tag{2.1.11}$$

The proof is now completed by applying Lemma 2.1.3. □

Combining Lemma 2.1.1 and Lemma 2.1.4 we prove Sanov's theorem for the finite alphabet Σ, our first large deviations principle. Later on, in Sections ?? and ?? we shall extend this result to an appropriate class of topological spaces $M_1(\Sigma)$ covering in particular the important case of $\Sigma = \mathbb{R}^d$.

Sanov's theorem deals with the sequence of laws governing the random empirical measure $L_n^{\mathbf{X}}$ which is a *random element* of the probability simplex $M_1(\Sigma)$ when the underlying independent random variables $X_1, \ldots, X_n$ are identically distributed according to the law μ.

Theorem 2.1.1 (Sanov's theorem) *For any set Γ of probability vectors in $M_1(\Sigma) \subset \mathbb{R}^{|\Sigma|}$*

$$-\inf_{\nu \in \Gamma^o} H(\nu|\mu) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) \leq -\inf_{\nu \in \Gamma} H(\nu|\mu)\,, \tag{2.1.12}$$

where Γ^o is the interior of Γ (as a set in $M_1(\Sigma) \subset \mathbb{R}^{|\Sigma|}$).

Remark: Comparing (2.1.12) and the definition (1.1.13) we conclude that Sanov's theorem states that the family of laws $\mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \cdot)$ satisfies a large deviations principle with the rate function $H(\cdot|\mu)$. Further, in the particular case of finite alphabet Σ covered here the upper bound holds for any set Γ (i.e., there is no need for a closure operation). For few other improvements over (2.1.12) which are specific to this case see exercises 2.1.1, 2.1.2 and 2.1.3. Note however that there are closed sets for which the upper and lower bounds of (2.1.12) are far apart (see exercise 2.1.5). Moreover, there are closed sets for which the limit of $\frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma)$ does not exist at all (see exercise 2.1.4).

Proof: We begin by deducing from Lemma 2.1.4 upper and lower bounds for any finite n. The upper bound

$$\text{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) = \sum_{\nu \in \Gamma \cap \mathcal{L}_n} \text{Prob}_\mu(L_n^{\mathbf{X}} = \nu) \leq \sum_{\nu \in \Gamma \cap \mathcal{L}_n} e^{-nH(\nu|\mu)} \leq |\Gamma \cap \mathcal{L}_n| e^{-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu)}$$

$$\leq (n+1)^{|\Sigma|} e^{-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu)} . \tag{2.1.13}$$

follows by the union of events bound and the upper bound of Lemma 2.1.4. The accompanying lower bound is

$$\text{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) = \sum_{\nu \in \Gamma \cap \mathcal{L}_n} \text{Prob}_\mu(L_n^{\mathbf{X}} = \nu) \geq \sum_{\nu \in \Gamma \cap \mathcal{L}_n} (n+1)^{-|\Sigma|} e^{-nH(\nu|\mu)} \geq$$

$$\geq (n+1)^{-|\Sigma|} e^{-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu)} . \tag{2.1.14}$$

As $\lim_{n\to\infty} \frac{1}{n} \log(n+1)^{|\Sigma|} = 0$ the normalized logarithmic limits of (2.1.13) and (2.1.14) yield

$$\limsup_{n\to\infty} \frac{1}{n} \log \text{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) = -\liminf_{n\to\infty} \{ \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu) \} \tag{2.1.15}$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log \text{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) = -\limsup_{n\to\infty} \{ \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu) \} . \tag{2.1.16}$$

The upper bound of (2.1.12) follows since $\Gamma \cap \mathcal{L}_n \subset \Gamma$ for any n.

Turning now to the lower bound of (2.1.12), fix an arbitrary point $\nu \in \Gamma^o$. The set $B_{\nu,\delta}^V \triangleq \{\nu' : d_V(\nu,\nu') < \delta\}$ must be contained in Γ for all $\delta > 0$ sufficiently small (as ν is in the interior of Γ). Thus, by part (b) of Lemma 2.1.1 (see (2.1.2)) there exists a sequence $\nu_n \in \Gamma \cap \mathcal{L}_n$ such that $\nu_n \to \nu$ as $n \to \infty$. As already observed, Σ may be identified with Σ_μ without loss of generality. Then $H(\cdot|\mu)$ is a continuous function and therefore,

$$-\limsup_{n\to\infty} \inf_{\nu' \in \Gamma \cap \mathcal{L}_n} H(\nu'|\mu) \geq -\lim_{n\to\infty} H(\nu_n|\mu) = -H(\nu|\mu) . \tag{2.1.17}$$

The lower bound of (2.1.12) follows by taking the infimum over $\nu \in \Gamma^o$ in (2.1.17). □

Exercises:

2.1.1 Prove that for any open set Γ

$$-\lim_{n\to\infty} \{ \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu|\mu) \} = \lim_{n\to\infty} \frac{1}{n} \log \text{Prob}_\mu (L_n^{\mathbf{X}} \in \Gamma) = -\inf_{\nu \in \Gamma} H(\nu|\mu) \triangleq -I_\Gamma . \tag{2.1.18}$$

2.1.2 (a). Extend the conclusions of exercise 2.1.1 to any set $\Gamma \subseteq M_1(\Sigma_\mu)$ which is contained in the closure of its interior.

(b). Prove that for any such set $I_\Gamma < \infty$ and $I_\Gamma = H(\nu^*|\mu)$ for some $\nu^* \in \overline{\Gamma^o}$.

Hint: Use the continuity of $H(\cdot|\mu)$ on $M_1(\Sigma_\mu)$ and the compactness of this set.

2.1.3 Assume that Γ is a convex subset of $M_1(\Sigma_\mu)$ of non-empty interior. Prove that all the conclusions of exercise 2.1.2 apply. Moreover, prove that here the point $\nu^* \in \overline{\Gamma^o}$ for which $I_\Gamma = H(\nu^*|\mu)$ is unique.

Hint: Show that for any $\nu \in \Gamma$ and any $\nu' \in \Gamma^o$ the line segment connecting ν and ν' is part of Γ^o. Deduce that $\Gamma \subset \overline{\Gamma^o}$ which is a closed convex set. Then prove that $H(\cdot|\mu)$ is a strictly convex function in $M_1(\Sigma_\mu)$.

2.1.4 Find a closed set Γ for which both of the limits in (2.1.18) do not exist.

Hint: Any set $\Gamma = \{\nu\}$ consisting of one probability vector $\nu \in M_1(\Sigma_\mu)$ which also belongs to $\mathcal{L}_n$ (for some n) shall do.

2.1.5 Find a closed set Γ such that $\Gamma = \overline{\Gamma^o}$ and $\inf_{\nu\in\Gamma} H(\nu|\mu) < \infty$ while $\inf_{\nu\in\Gamma^o} H(\nu|\mu) = \infty$.

Hint: For this construction you need $\Sigma_\mu \neq \Sigma$. Try $\Gamma = \{\nu : \nu(a_2) + \sqrt{\nu(a_1)} \geq 1\}$ and $\mu(a_1) = 0$ (while $|\Sigma| = 3$).

2.1.2 Cramer's theorem for finite alphabets in $\mathbb{R}^1$

As an application of Sanov's theorem, we bring a proof of a version of Cramer's theorem about the large deviations of the empirical mean of i.i.d random variables. For that purpose we further assume throughout this section that $\Sigma = \Sigma_\mu$ is a finite subset of $\mathbb{R}^1$ and let $\hat{S}_n \triangleq \frac{1}{n}\sum_{j=1}^n X_j$ denote the resulting sequence of empirical means (where $X_j \in \Sigma$ as in Section 2.1.1 above). Cramer's theorem deals with the large deviations principle satisfied by the family of laws governing the real valued random variables $\hat{S}_n$. Sections 2.2 and 2.3 are devoted to successive generalizations of this result to $\Sigma = \mathbb{R}^1$ (Section 2.2) and to weakly dependent random vectors in $\mathbb{R}^d$ (Section 2.3).

Note that in the case considered here ($|\Sigma| < \infty$), the random variables $\hat{S}_n$ assume values in the compact interval $K \triangleq [\min_{i=1}^{|\Sigma|}\{a_i\}, \max_{i=1}^{|\Sigma|}\{a_i\}]$. Moreover,

$\hat{S}_n = \sum_{i=1}^{|\Sigma|} L_n^{\mathbf{X}}(a_i)a_i =< L_n^{\mathbf{X}}, \mathbf{a} >$ where $\mathbf{a} \stackrel{\Delta}{=} (a_1, \ldots, a_{|\Sigma|})$. Therefore,

$$\hat{S}_n \in A \quad \Leftrightarrow \quad L_n^{\mathbf{X}} \in \{\nu : < \nu, \mathbf{a} > \in A\} \stackrel{\Delta}{=} \Gamma \tag{2.1.19}$$

for any $A \subset K$ and any integer n. Thus, the following version of Cramer's theorem is a direct consequence of Theorem 2.1.1.

Theorem 2.1.2 (Cramer's theorem for finite subsets of $\mathbb{R}^1$)

(a). For any set $A \subset K$

$$-\inf_{x \in A^o} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mathrm{Prob}_\mu(\hat{S}_n \in A) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mathrm{Prob}_\mu(\hat{S}_n \in A) \leq -\inf_{x \in A} I(x) \,, \tag{2.1.20}$$

where A^o is the interior of A (as a set in $\mathbb{R}^1$) and $I(x) \stackrel{\Delta}{=} \inf_{<\nu, \mathbf{a}>=x} H(\nu|\mu)$.

(b). The continuous rate function $I(x)$ (for $x \in K$) also satisfies

$$I(x) = \sup_{\lambda \in \mathbb{R}^1} \{\lambda x - \Lambda(\lambda)\} \,, \tag{2.1.21}$$

where

$$\Lambda(\lambda) = \log \sum_{i=1}^{|\Sigma|} \mu(a_i) e^{\lambda a_i} \,.$$

Remark: Since the rate function is continuous (on K) it follows from (2.1.20) that whenever $A \subset \overline{A^o} \subset K$

$$\lim_{n \to \infty} \frac{1}{n} \log \mathrm{Prob}_\mu(\hat{S}_n \in A) = -\inf_{x \in A} I(x) \,.$$

Proof:

(a). The bounds of (2.1.20) are simply the bounds of (2.1.12) for the set Γ defined in (2.1.19) (note that $\{\nu :< \nu, \mathbf{a} >\in A^o\} \subset \Gamma^o$).

(b). Observe that by Jensen's inequality

$$\Lambda(\lambda) = \log \sum_{i=1}^{|\Sigma|} \mu(a_i) e^{\lambda a_i} \geq \sum_{i=1}^{|\Sigma|} \nu(a_i) \log \frac{\mu(a_i) e^{\lambda a_i}}{\nu(a_i)} = \lambda < \nu, \mathbf{a} > - H(\nu|\mu) \,, \tag{2.1.22}$$

for any $\nu \in M_1(\Sigma)$ and any $\lambda \in \mathbb{R}^1$, with equality for $\nu_\lambda(a_i) \stackrel{\Delta}{=} \mu(a_i) e^{\lambda a_i - \Lambda(\lambda)}$.

The function $\Lambda(\lambda)$ is differentiable and strictly convex, implying that $\Lambda'(\lambda)$ is strictly increasing. Moreover, $K = [\inf_\lambda \Lambda'(\lambda) \,, \sup_\lambda \Lambda'(\lambda)]$. Therefore, for any $x \in K^o$ there exists a unique $\lambda^{(x)} \in \mathbb{R}^1$

such that $x - \Lambda'(\lambda^{(x)}) = 0$. As $< \nu_{\lambda^{(x)}}, a >= \Lambda'(\lambda^{(x)}) = x$, the application of (2.1.22) for $\lambda^{(x)}$ yields

$$I(x) = \inf_{<\nu,a>=x} H(\nu|\mu) = H(\nu_{\lambda^{(x)}}|\mu) = \lambda^{(x)}x - \Lambda(\lambda^{(x)}) = \sup_{\lambda\in\mathbb{R}^1} \{\lambda x - \Lambda(\lambda)\}\,, \qquad (2.1.23)$$

i.e., (2.1.21) holds for any $x \in K^o$. The continuity of $I(x)$ (for $x \in K$), which is a direct consequence of the continuity of $H(\cdot|\mu)$, implies that (2.1.21) holds also at the boundaries of K. □

It is interesting to note that by deriving Cramer's theorem from Sanov's theorem, we have followed a pattern which will be useful in the sequel and which is referred to as the 'contraction principle' (cf Section ??). Indeed, based on large deviations result in the "big" space ($M_1(\Sigma)$), we obtained a large deviations result on a "smaller" space which is obtained from the original space by a continuous map ("contraction"). In the particular case in hand, however, an alternative proof of the finite alphabet Sanov's theorem follows directly from Cramer's theorem in $\mathbb{R}^{|\Sigma|}$ and is presented in section 2.3.

Exercises:

2.1.6 Construct an example for which the limit of $\frac{1}{n}\log \mathrm{Prob}_\mu(\hat{S}_n = x)$ as $n \to \infty$ does not exist. **Hint:** Note that for $|\Sigma| = 2$ the empirical mean $(\hat{S}_n)$ uniquely determines the empirical measure (L_n^X). Relay on this observation and exercise 2.1.4 to construct the desired example.

2.1.7 (a). Prove that $I(x) = 0$ if and only if $x = \overline{x} \stackrel{\Delta}{=} E_\mu(X_1)$. Explain why this should have been anticipated in view of the weak law of large numbers (which states that $\hat{S}_n \to \overline{x}$ in probability, as $n \to \infty$).
(b). Check that $H(\nu|\mu) = 0$ if and only if $\nu = \mu$ and interpret this result.

2.1.8 Guess the value of $\lim_{n\to\infty} \mathrm{Prob}_\mu(X_1 = a_i|\hat{S}_n \geq q)$ for $\overline{x} < q \in K^o$ and try to justify your guess (at least heuristically).

2.1.9 Extend Theorem 2.1.2 to Σ which is a finite subset of $\mathbb{R}^d$. In particular, determine the shape of the set K and find the appropriate extension of the formula (2.1.21).

2.1.3 Large deviations for sampling without replacements

The method of types is useful also for studying the large deviations of the empirical measure of a sequence of dependent random variables. For example, consider the following set up of sampling

without replacements which is often encountered in statistical problems. Out of an initial deterministic pool of M items $\mathbf{x} \stackrel{\Delta}{=} (x_1, \ldots, x_M)$, an n-tuple $\mathbf{X} \stackrel{\Delta}{=} (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ is sampled without replacement, namely, $i_1 \neq i_2 \neq i_3 \neq \ldots \neq i_n$ and all $\begin{pmatrix} M \\ n \end{pmatrix}$ choices of $i_1 \neq i_2 \ldots \neq i_n \in \{1, \ldots, M\}$ are equally likely.

Suppose now that $x_1^{(M)}, \ldots, x_M^{(M)}$ are elements of the same finite set $\Sigma = \{a_1, \ldots, a_{|\Sigma|}\}$ and that as $M \to \infty$ the deterministic relative frequency vectors $L_M^{\mathbf{x}} = \{L_M^{\mathbf{x}}(a_1), \ldots, L_M^{\mathbf{x}}(a_{|\Sigma|})\}$ converge (point wise) to $\mu \in M_1(\Sigma)$. Recall that

$$L_M^{\mathbf{x}}(a_i) = \frac{1}{M} \sum_{j=1}^{M} 1_{x_j^{(M)}=a_i} \stackrel{\Delta}{=} \frac{1}{M} M(a_i|\,\mathbf{x})\,, \quad i = 1, \ldots, |\Sigma| \ . \tag{2.1.24}$$

Suppose further that $\mathbf{X}$ is a random vector obtained by the sampling without replacement of n out of M elements as described above. We investigate the large deviations principle governing the random empirical measure $L_n^{\mathbf{X}}$ associated with the vector $\mathbf{X}$. Note that $L_n^{\mathbf{X}}$ belongs to the set $\mathcal{L}_n$ whose size grows polynomially in n (see Lemma 2.1.1). In particular we aim at deriving the analog of Theorem 2.1.1 when $\lim_{M\to\infty} (\frac{n}{M}) = \beta \in (0,1)$. For that purpose define the following candidate rate function

$$I_{\beta,\mu}(\nu) \stackrel{\Delta}{=} \begin{cases} H(\nu|\mu) + \frac{1-\beta}{\beta} H\left(\frac{\mu-\beta\nu}{1-\beta}\middle|\mu\right) & \text{if } \mu_i \geq \beta\nu_i \text{ for all } i \\ \infty & \text{otherwise} \end{cases} \tag{2.1.25}$$

By basic combinatorics and an application of Lemma 2.1.3 we obtain the following estimates of large deviations probabilities for $L_n^{\mathbf{X}}$.

Lemma 2.1.5 *For any probability vector $\nu \in \mathcal{L}_n$*

(a). If $I_{\frac{n}{M}, L_M^{\mathbf{x}}}(\nu) < \infty$ then

$$\left|\frac{1}{n} \log \text{Prob}(L_n^{\mathbf{X}} = \nu) + I_{\frac{n}{M}, L_M^{\mathbf{x}}}(\nu)\right| \leq 2(|\Sigma| + 1) \left(\frac{\log(M+1)}{n}\right) \tag{2.1.26}$$

(b). If $I_{\frac{n}{M}, L_M^{\mathbf{x}}}(\nu) = \infty$ then

$$\text{Prob}(L_n^{\mathbf{X}} = \nu) = 0 \tag{2.1.27}$$

Proof: (a). In the sampling without replacement procedure the probability of the event $\{L_n^{\mathbf{X}} = \nu\}$ for $\nu \in \mathcal{L}_n$ is exactly the number of n–tuples $i_1 \neq i_2 \cdots \neq i_n$ resulting with type ν divided by the

overall number of n–tuples, that is

$$\text{Prob}(L_n^{\mathbf{X}} = \nu) = \frac{\prod_{i=1}^{|\Sigma|} \binom{M\, L_M^{\mathbf{x}}(a_i)}{n\,\nu(a_i)}}{\binom{M}{n}}. \tag{2.1.28}$$

An application of Lemma 2.1.3 for $|\Sigma| = 2$ (where $|T(\nu)| = \binom{n}{k}$ when $\nu(a_1) = \frac{k}{n}$, $\nu(a_2) = 1 - \frac{k}{n}$), results with the following estimate

$$\max_{0 \le k \le n} \left| \log \binom{n}{k} - nH\left(\frac{k}{n}\right) \right| \le 2\log(n+1)\,, \tag{2.1.29}$$

where

$$H(p) \stackrel{\Delta}{=} -p\log p - (1-p)\log(1-p)\,.$$

Alternatively, (2.1.29) follows by Stirling's formula (see [12], pg 48).

Combining the exact expression (2.1.28) and the bound of (2.1.29) results with

$$\left| \frac{1}{n} \log \text{Prob}\,(L_n^{\mathbf{X}} = \nu) - \sum_{i=1}^{|\Sigma|} \frac{M\, L_M^{\mathbf{x}}(a_i)}{n}\, H\left(\frac{n\nu(a_i)}{M L_M^{\mathbf{x}}(a_i)}\right) + \frac{M}{n} H\left(\frac{n}{M}\right) \right| \le 2(|\Sigma|+1)\left(\frac{\log(M+1)}{n}\right) \tag{2.1.30}$$

The inequality (2.1.26) follows when rearranging the left side of (2.1.30).

(b). Note that $I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu) = \infty$ if and only if $n\nu(a_i) > M L_M^{\mathbf{x}}(a_i) = M(a_i|\mathbf{x})$ for some $a_i \in \Sigma$. It is then impossible in sampling without replacement to have $L_n^{\mathbf{X}} = \nu$, as $nL_n^{\mathbf{X}}(a_i) = N(a_i|\mathbf{X}) \le M(a_i|\mathbf{x})$ for any $a_i \in \Sigma$. □

Following the proof of Theorem 2.1.1 the above point estimates result with the analogs of (2.1.15) and (2.1.16), namely:

Corollary 2.1.1

$$\limsup_{M\to\infty} \frac{1}{n} \log \text{Prob}(L_n^{\mathbf{X}} \in \Gamma) = -\liminf_{M\to\infty} \{ \inf_{\nu\in\Gamma\cap\mathcal{L}_n} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu) \} \tag{2.1.31}$$

and

$$\liminf_{M\to\infty} \frac{1}{n} \log \text{Prob}(L_n^{\mathbf{X}} \in \Gamma) = -\limsup_{M\to\infty} \{ \inf_{\nu\in\Gamma\cap\mathcal{L}_n} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu) \}\,. \tag{2.1.32}$$

The proof of Corollary 2.1.1 is left as exercise 2.1.10.

The following theorem is the desired analog of Theorem 2.1.1.

Theorem 2.1.3 *The good rate function $I_{\beta,\mu}(\nu)$ controls the large deviations principle related to the random empirical measures $L_n^{\mathbf{X}}$ (as elements of $M_1(\Sigma)$). Specifically, for any set Γ of probability vectors in $M_1(\Sigma) \subset \mathbb{R}^{|\Sigma|}$,*

$$-\inf_{\nu\in\Gamma^o} I_{\beta,\mu}(\nu) \leq \liminf_{M\to\infty} \frac{1}{n}\log \text{Prob}(L_n^{\mathbf{X}} \in \Gamma) \leq \limsup_{M\to\infty} \frac{1}{n}\log \text{Prob}(L_n^{\mathbf{X}} \in \Gamma) \leq -\inf_{\nu\in\bar{\Gamma}} I_{\beta,\mu}(\nu)\ , \tag{2.1.33}$$

where $\beta = \lim_{M\to\infty}(\frac{n}{M}) \in (0,1)$ and $L_M^{\mathbf{X}}$ converges to μ as $M \to \infty$.

Remark: Note that the upper bound of (2.1.33) is weaker than the upper bound of (2.1.12). Also, see exercise 2.1.11 for examples of sets for which the above lower and upper bounds coincide.

Proof: As already noted, $H(\cdot|\mu)$ is a rate function on $M_1(\Sigma)$, for any fixed μ (namely, it is lower semicontinuous and non-negative). Moreover, the set $\{\nu : (\mu - \beta\nu)/(1-\beta) \in M_1(\Sigma)\}$ is a closed subset of $M_1(\Sigma)$ and therefore, $I_{\beta,\mu}(\cdot)$ is also a rate function for any fixed $\beta \in (0,1)$ and any fixed $\mu \in M_1(\Sigma)$. Since $|\Sigma| < \infty$, the probability simplex $M_1(\Sigma)$ is a compact set and thus any rate function on $M_1(\Sigma)$ is a good rate function.

The first step in deriving the bounds of (2.1.33) is to prove that $I_{\beta,\mu}(\nu)$ is a lower semicontinuous function over $(0,1) \times M_1(\Sigma) \times M_1(\Sigma)$ (jointly in β, μ and ν) and is strictly continuous along sequences $\{\beta_n, \mu_n, \nu_n\}$ where $I_{\beta_n,\mu_n}(\nu_n) < \infty$. For that purpose, fix a sequence $\{\beta_n, \mu_n, \nu_n\}$ such that $\beta_n \to \beta \in (0,1)$, $\nu_n \to \nu$ and $\mu_n \to \mu$. The lower semicontinuity is trivial along any subsequence $\{n_k\}_{k=1}^{\infty}$ for which $I_{\beta_{n_k},\mu_{n_k}}(\nu_{n_k}) = \infty$. Thus, without loss of generality assume that $I_{\beta_n,\mu_n}(\nu_n) < \infty$ for all n large enough, namely that $\mu_n(a_i) \geq \beta_n\nu_n(a_i)$ for any $a_i \in \Sigma$. Then, by rearranging (2.1.25)

$$I_{\beta_n,\mu_n}(\nu_n) = \frac{1}{\beta_n}H(\mu_n) - H(\nu_n) - \frac{1-\beta_n}{\beta_n}H(\frac{\mu_n - \beta_n\nu_n}{1-\beta_n})\ .$$

As β_n are bounded away from 0 and 1 for large n, it follows from the above expression that $I_{\beta_n,\mu_n}(\nu_n) \longrightarrow I_{\beta,\mu}(\nu)$ since the entropy function $H(\cdot)$ is strictly continuous over $M_1(\Sigma)$ (for $|\Sigma| < \infty$).

We turn now to prove the upper bound of (2.1.33). We first deduce from (2.1.31) that for some

infinite subsequence M_k, there exist $\nu_k \in \Gamma$ such that

$$\limsup_{M\to\infty} \frac{1}{n} \log \text{Prob}(L_n^{\mathbf{X}} \in \Gamma) = -\lim_{k\to\infty} I_{\frac{n}{M_k},L_{M_k}^{\mathbf{X}}}(\nu_k) \stackrel{\Delta}{=} I^* \ . \tag{2.1.34}$$

Moreover, the sequence $\{\nu_k\}$ has a limit point ν^* in the *compact* set $\overline{\Gamma}$. Passing to a convergent subsequence, the lower semicontinuity of I jointly in β, μ and ν implies that

$$I^* \leq -I_{\beta,\mu}(\nu^*) \leq -\inf_{\nu\in\overline{\Gamma}} I_{\beta,\mu}(\nu) \ . \tag{2.1.35}$$

The proof of the upper bound is complete in view of (2.1.34) and (2.1.35).

In view of (2.1.32) it suffices for the lower bound of (2.1.33) to prove that

$$-\limsup_{M\to\infty} \{ \inf_{\nu'\in\Gamma\cap\mathcal{L}_n} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu') \} \geq -I_{\beta,\mu}(\nu) \tag{2.1.36}$$

for any $\nu \in \Gamma^\circ$ where $I_{\beta,\mu}(\nu) < \infty$ (see (1.1.7) in Section 1.1 for this formulation of the large deviations lower bound). Fix such a point ν, i.e., such that

$$\mu(a_i) \geq \beta\nu(a_i) \quad \forall a_i \in \Sigma \ . \tag{2.1.37}$$

In particular, $\Sigma_\nu \subseteq \Sigma_\mu$ and as $\nu \in \Gamma^\circ$ as well, ν belongs to the relative interior of Γ in $M_1(\Sigma_\mu)$. Further, (2.1.36) follows from

$$-\limsup_{M\to\infty} \{ \inf_{\nu'\in\Gamma\cap M_1(\Sigma_\mu)\cap\mathcal{L}_n} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu') \} \geq -I_{\beta,\mu}(\nu) \ .$$

Therefore, one may assume without loss of generality that $\Sigma = \Sigma_\mu$. Now, since $\beta < 1$ there always exists a sequence of probability vectors $\nu_k \in \Gamma^\circ$ which converges to ν and for which *all* the inequalities in (2.1.37) are *strict*. As $I_{\beta,\mu}(\cdot)$ is continuous along any such sequence, we deduce that it suffices to prove (2.1.36) for $\nu \in \Gamma^\circ$ for which

$$\min_{a_i\in\Sigma}\{\mu(a_i) - \beta\nu(a_i)\} > 0 \ . \tag{2.1.38}$$

By the same argument as in the proof of Theorem 2.1.1 there exists a sequence $\nu_n \in \Gamma \cap \mathcal{L}_n$ such that $\nu_n \to \nu$ as $n \to \infty$. Now, since $L_M^{\mathbf{X}} \to \mu$ and $\frac{n}{M} \to \beta$ the *strict* inequality (2.1.38) implies that for all M sufficiently large

$$\min_{a_i\in\Sigma}\{L_M^{\mathbf{X}}(a_i) - \frac{n}{M}\nu_n(a_i)\} \geq 0$$

and then $I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu_n) < \infty$ (note also that $\frac{n}{M}$ is eventually bounded away from 0 and 1). Therefore, by the strict continuity of I along such sequences

$$I_{\beta,\mu}(\nu) = \lim_{M\to\infty} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu_n) \ .$$

Since $\nu_n \in \Gamma \cap \mathcal{L}_n$,

$$\limsup_{M\to\infty} \{ \inf_{\nu\in\Gamma\cap\mathcal{L}_n} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu) \} \leq \lim_{M\to\infty} I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu_n)$$

and the inequality (2.1.36) now follows. □

Exercises:

2.1.10 Prove Corollary 2.1.1.

2.1.11 Let $I_\Gamma \stackrel{\Delta}{=} \inf_{\nu\in\Gamma^\circ} I_{\beta,\mu}(\nu)$. Prove that when $I_\Gamma < \infty$ and Γ is included in the closure of its interior, then

$$I_\Gamma = \lim_{M\to\infty} \frac{1}{n} \log \mathrm{Prob}(L_n^{\mathbf{X}} \in \Gamma) \ .$$

Hint: See exercise 2.1.2 and use the continuity of $I_{\beta,\mu}(\cdot)$ within its level sets.

2.1.12 Prove that the rate function $I_{\beta,\mu}(\cdot)$ is a convex function.

2.1.13 Prove that the large deviations principle of Theorem 2.1.3 holds for $\beta = 0$ with the good rate function $I_{0,\mu}(\cdot) \stackrel{\Delta}{=} H(\cdot|\mu)$. provided that $n - \infty$.
Hint: First prove that the left side of (2.1.30) goes to zero as long as $n \to \infty$ (i.e., even when $\left(\frac{\log(M+1)}{n}\right) \to \infty$). Then prove the lower semicontinuity of $I_{\beta,\mu}(\nu)$ at $\beta = 0$ and use it to derive the upper bound. Finally, it suffices to prove the lower bound when $\Sigma_\nu \subseteq \Sigma_\mu$ and so the sequence $\nu_n \in M_1(\Sigma_\mu) \cap \Gamma \cap \mathcal{L}_n$ will converge to ν while eventually $I_{\frac{n}{M},L_M^{\mathbf{X}}}(\nu_n) < \infty$.

2.2 Cramer's Theorem in $\mathbb{R}^1$

Cramer's theorem about the large deviations associated with the empirical means of i.i.d. random variables is presented in Section 2.1.2 as an application of the method of types. However, the method of types is limited in its scope to finite alphabets. Moreover, it neither explains why a large deviations principle is satisfied nor predicts which rate functions one should expect in similar

situations. In this section, we start pursuing a different route aiming at proving the analog of Theorem 2.1.2 when the underlying alphabet Σ is $\mathbb{R}^1$. The approach outlined here is more amenable for generalizations and better illustrates the main ingredients involved in proving a typical large deviations principle.

Consider the empirical means $\hat{S}_n \stackrel{\Delta}{=} \frac{1}{n} \sum_{j=1}^{n} X_j$, where the **real valued** random variables $X_1, \ldots, X_n, \ldots$ are independent and identically distributed according to the marginal probability law $\mu \in M_1(\mathbb{R})$. Let μ_n denote the law of $\hat{S}_n$ and $\overline{x} \stackrel{\Delta}{=} E_\mu[X_1]$ denotes the true underlying mean. If we further assume that $\overline{x} < \infty$ and $E[|X_1 - \overline{x}|^2] < \infty$ then $\hat{S}_n \underset{n\to\infty}{\overset{Prob}{\longrightarrow}} \overline{x}$ since

$$E[(\hat{S}_n - \overline{x})^2] = \frac{1}{n^2} \sum_{j=1}^{n} E[|X_j - \overline{x}|^2] = \frac{1}{n} E[|X_1 - \overline{x}|^2] \underset{n\to\infty}{\longrightarrow} 0 \,. \tag{2.2.39}$$

Actually, the finite variance condition is not needed for the convergence in probability but we will not care about that here. Because of (2.2.39), $\mu_n(A) \underset{n\to\infty}{\longrightarrow} 0$ for any set A such that $\overline{x} \notin \overline{A}$. Cramer's theorem characterizes the logarithmic rate of this convergence by the (rate) function

$$\Lambda^*(x) \stackrel{\Delta}{=} \sup_{\lambda \in \mathbb{R}^1} [\lambda x - \Lambda(\lambda)] \,, \tag{2.2.40}$$

where

$$\Lambda(\lambda) \stackrel{\Delta}{=} \log M(\lambda) \stackrel{\Delta}{=} \log E[e^{\lambda X_1}] \,, \tag{2.2.41}$$

is also called the **logarithmic moment generating function.** Note that while $\Lambda(\lambda) > -\infty$ it is possible to have $\Lambda(\lambda) = \infty$. Let $\mathcal{D}_\Lambda \stackrel{\Delta}{=} \{\lambda : \Lambda(\lambda) < \infty\}$ and $\mathcal{D}_{\Lambda^*} \stackrel{\Delta}{=} \{x : \Lambda^*(x) < \infty\}$.

The rate function $\Lambda^*(x)$ of (2.2.40) is called the *Legendre transform* of $\Lambda(\lambda)$. Some of the properties of this function (and of $\Lambda(\lambda)$) which are useful when proving Cramer's theorem are summarized in the following lemma whose proof is deferred to the end of this section. The exact definition of Legendre transform and its properties for more general vector spaces are presented in Sections 2.3 (for $\mathbb{R}^d$) and ?? (for a general class of metric spaces).

Lemma 2.2.1

(a). Λ *is a convex function and* Λ^* *is a convex rate function.*

(b). $\Lambda^*(\overline{x}) = 0$ *and for* $x \in [\overline{x}, \infty)$,

$$\Lambda^*(x) = \sup_{\lambda \geq 0} [\lambda x - \Lambda(\lambda)] \,, \tag{2.2.42}$$

is nondecreasing. Similarly, for $x \in (-\infty, \overline{x}]$

$$\Lambda^*(x) = \sup_{\lambda \leq 0}[\lambda x - \Lambda(\lambda)] \tag{2.2.43}$$

and is nonincreasing.

(c). For any $\eta \in \mathcal{D}_\Lambda^o$

$$\Lambda'(\eta) = \frac{1}{M(\eta)} E[X_1 e^{\eta X_1}] \tag{2.2.44}$$

and

$$\Lambda'(\eta) = y \quad \Longrightarrow \quad \Lambda^*(y) = \eta y - \Lambda(\eta) \tag{2.2.45}$$

The following definition is needed for the precise statement of Cramer's theorem.

Definition 2.2.1 *The logarithmic moment generating function* $\Lambda(\cdot)$ *is* **steep** *if* $\liminf_{n\to\infty} |\Lambda'(\lambda_n)| = \infty$ *for any sequence* $\lambda_n \in \mathcal{D}_\Lambda^o$ *which converges to* $\lambda \in \mathcal{D}_\Lambda \setminus \mathcal{D}_\Lambda^o$.

A weak version of Cramer's theorem is proved below, establishing the large deviations principle in $\mathbb{R}^1$ when $\Lambda(\cdot)$ is a steep function and $0 \in \mathcal{D}_\Lambda^o$. For example, both of these conditions hold when $\mathcal{D}_\Lambda = \mathbb{R}^1$. It is possible to establish the large deviations principle in $\mathbb{R}^1$ with no condition on $\Lambda(\lambda)$ by adopting a more sophisticated proof based on sub-additivity and convexity arguments. This is pursued in much generality in Sections ?? and ?? where the exact consequences for the special case discussed here are given in exercise ??.

Theorem 2.2.1 (Cramer)

(a). The large deviations upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq - \inf_{x \in \overline{A}} \Lambda^*(x) , \tag{2.2.46}$$

holds for any measurable $A \subset \mathbb{R}^1$.

(b). For any open set $G \subset \mathbb{R}^1$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x \in G \cap \mathcal{F}} \Lambda^*(x) , \tag{2.2.47}$$

where $\mathcal{F} \stackrel{\Delta}{=} \{x : x = \Lambda'(\lambda) \ \ \textit{for some} \ \ \lambda \in \mathcal{D}_\Lambda^o\}$.

(c). If $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$ *then the family of probability measures* μ_n *satisfies the large deviations principle*

with the rate function $\Lambda^*(x)$. *In particular,* μ_n *satisfies the large deviations principle if* $\mathcal{D}_\Lambda = \mathbb{R}^1$.
(d). If $0 \in \mathcal{D}_\Lambda^o$ *and* Λ *is a steep function then the family of probability measures* μ_n *satisfies the large deviations principle with the good rate function* $\Lambda^*(x)$.

Remark: The Legendre transform $\Lambda^*(x)$ is a natural candidate for the rate function under a very general situation where the upper bound (2.2.46) holds (for more on this issue, see Sections 2.3 and ??). For that reason, convexity plays an important role in large deviations as seen throughout Chapters ??-??.

Proof:
(a). If $\overline{x} \in \overline{A}$ then the upper bound (2.2.46) trivially holds (as $\Lambda^*(\overline{x}) = 0$). Otherwise, let (x_-, x_+) be the largest interval around $\overline{x}$ such that $(x_-, x_+) \cap \overline{A} = \emptyset$. Since

$$\overline{A} \subseteq (-\infty, x_-] \cup [x_+, \infty) \stackrel{\Delta}{=} A_- \cup A_+ ,$$

it suffices to prove that

$$\mu_n(A_+) \leq e^{-n\Lambda^*(x_+)} \tag{2.2.48}$$

and

$$\mu_n(A_-) \leq e^{-n\Lambda^*(x_-)} \tag{2.2.49}$$

in order to prove the upper bound. Indeed, then

$$\mu_n(A) \leq \mu_n(A_+) + \mu_n(A_-) \leq e^{-n\Lambda^*(x_+)} + e^{-n\Lambda^*(x_-)} \leq 2e^{-n\min\{\Lambda^*(x_-), \Lambda^*(x_+)\}} .$$

By part (b) of Lemma 2.2.1

$$\min\{\Lambda^*(x_-), \Lambda^*(x_+)\} = \inf_{x \in A} \Lambda^*(x)$$

and the proof of the upper bound follows.

Returning now to prove (2.2.48) and (2.2.49), these are consequences of a special form of Markov's inequality. Clearly, for any $\lambda \geq 0$,

$$\begin{aligned} \mu_n(A_+) &\leq \int_{A_+} e^{n\lambda(z-x_+)}\, d\mu_n(z) &&\leq e^{-n\lambda x_+} E(e^{\lambda n \hat{S}_n}) \\ &= e^{-n\lambda x_+} \textstyle\prod_{i=1}^n E(e^{\lambda X_i}) &&= e^{-n[\lambda x_+ - \Lambda(\lambda)]} \end{aligned} \tag{2.2.50}$$

Therefore,

$$\mu_n(A_+) \leq e^{-\sup_{\lambda \geq 0} n[\lambda x_+ - \Lambda(\lambda)]}$$

and (2.2.48) follows by (2.2.42) as $x_+ \in [\overline{x}, \infty)$. The proof of (2.2.49) is similar.

(b). The lower bound of (2.2.47) trivially holds when $G \cap \overline{\mathcal{F}}$ is an empty set. Moreover, since G is an open set it suffices to show that for any $x \in \overline{\mathcal{F}}$, and any open interval $B_{x,\delta}$ of center at x (and width 2δ)

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{x,\delta}) \geq -\Lambda^*(x) \,. \tag{2.2.51}$$

Consider first points $y \in \mathcal{F}$. Fix such a point $y = \Lambda'(\eta) \in \mathcal{F}$ and fix $\delta > 0$. Note that $\eta \in \mathcal{D}_\Lambda^o$ and define the associated probability measure (see [12])

$$\tilde{\mu}(dx) = e^{\eta x - \Lambda(\eta)} \mu(dx) \,.$$

Indeed $\tilde{\mu}$ is a probability measure as $\int_{\mathbb{R}^1} \tilde{\mu}(dx) = \frac{1}{M(\eta)} \int_{\mathbb{R}^1} e^{\eta x} \mu(dx) = 1$. Define accordingly $\tilde{\mu}_n$ to be the law governing $\hat{S}_n$ when X_i are i.i.d. random variables of law $\tilde{\mu}$. Note that by (2.2.44)

$$E_{\tilde{\mu}}(X_1) \stackrel{\Delta}{=} \frac{1}{M(\eta)} \int_{\mathbb{R}^1} x e^{\eta x} \mu(dx) = \Lambda'(\eta) = y \,.$$

Thus, the mean of X_i under $\tilde{\mu}$ equals y and by the weak law of large numbers,

$$\lim_{n\to\infty} \tilde{\mu}_n(B_{y,\delta}) = 1 \,. \tag{2.2.52}$$

Moreover,

$$\begin{aligned} \mu_n(B_{y,\delta}) &= \int_{|\hat{S}_n - y| < \delta} d\mu(x_1) \cdots d\mu(x_n) \geq e^{-n(\eta y + \delta|\eta|)} \int_{|\hat{S}_n - y| < \delta} e^{\eta \sum_{i=1}^n x_i} d\mu(x_1) \cdots d\mu(x_n) \\ &= \tilde{\mu}_n(B_{y,\delta}) e^{-n(\eta y + |\eta|\delta) + n\Lambda(\eta)} = \tilde{\mu}_n(B_{y,\delta}) e^{-n\Lambda^*(y)} e^{-n|\eta|\delta} \end{aligned} \tag{2.2.53}$$

where the last equality follows by (2.2.45). Therefore, by combining (2.2.52) and (2.2.53)

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{y,\delta}) \geq -\Lambda^*(y) - |\eta|\delta$$

Take now $\delta_m \to 0$, $\delta_m < \delta$. Then, from the above,

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{y,\delta}) &\geq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{y,\delta_m}) \\ &\geq -\Lambda^*(y) - |\eta|\delta_m \underset{m\to\infty}{\longrightarrow} -\Lambda^*(y) \end{aligned}$$

and the proof of (2.2.51) is complete for $y \in \mathcal{F}$.

The inequality (2.2.51) trivially holds when $x \notin \mathcal{D}_{\Lambda^*}$, while when $x \in \mathcal{D}_{\Lambda^*} \cap \overline{\mathcal{F}}$ there exist $y_r \in \mathcal{F} \subseteq \mathcal{D}_{\Lambda^*}$ such that $y_r \to x$. Since $\Lambda^*(\cdot)$ is a convex rate function, $\Lambda^*(y_r) \to \Lambda^*(x)$ and as $B_{x,\delta}$ is an open set, eventually $B_{y_r,\delta_r} \subseteq G$ for some $\delta_r > 0$. Finally, apply (2.2.51) for $y_r \in \mathcal{F}$ to obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{x,\delta}) \geq \lim_{r\to\infty} \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_{y_r,\delta_r}) \geq -\lim_{r\to\infty} \Lambda^*(y_r) = -\Lambda^*(x) .$$

(c+d). If $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$ then for any $G \subset \mathbb{R}^1$

$$-\inf_{x\in G\cap\overline{\mathcal{F}}} \Lambda^*(x) \geq -\inf_{x\in G\cap\mathcal{D}_{\Lambda^*}} \Lambda^*(x) = -\inf_{x\in G} \Lambda^*(x)$$

The large deviations principle now follows by combining (2.2.46) and (2.2.47). The claim is a direct consequence of parts (b) and (c) in Lemma 2.2.2 below. □

Remarks:

(a). The crucial step in the above proof of the upper bound is the exponential form of Markov's inequality combined with the independence assumption by which one can decompose the bounding exponent. For weakly dependent random variables a similar approach is to use the logarithmic limit of the left side of (2.2.50) instead of the logarithmic moment generating function for a single random variable. This is further explored in Section 2.3.

(b). The crucial step in the above proof of the lower bound is the exponential change of measure applied when defining the associated measure $\tilde{\mu}$. This is particularly well suited to problems where, even if the random variables involved are not directly independent, some form of underlying independence exists (e.g., when a Girsanov type formula can be used, as in Chapter ??). Unless coupled with some convexity, this argument may in general fail in the dependent case. For more about it, c.f. Section ??.

The following lemma lifts some of the mystery behind the condition $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$.

Lemma 2.2.2

(a). $\mathcal{F} \subseteq \mathcal{D}_{\Lambda^*}$ and both are intervals. Also $0 \in \mathcal{D}_\Lambda$ which is an interval.

(b). If $0 \in \mathcal{D}_\Lambda^o$ and Λ is a steep function then $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$.

(c). If $0 \in \mathcal{D}_\Lambda^o$ then Λ^* is a good rate function.

Remark: Both the steepness of Λ and the condition $0 \in \mathcal{D}_\Lambda^o$ are necessary for $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$ (for details see exercise 2.2.3).

Proof of lemma 2.2.1:

(a). The convexity of Λ follows by Holder's inequality, as for any $\theta \in [0,1]$

$$\Lambda(\theta\lambda_1+(1-\theta)\lambda_2) = \log E[(e^{\lambda_1 X_1})^{\theta}(e^{\lambda_2 X_1})^{(1-\theta)}] \leq \log\{E[e^{\lambda_1 X_1}]^{\theta} E[e^{\lambda_2 X_1}]^{(1-\theta)}\} = \theta\Lambda(\lambda_1)+(1-\theta)\Lambda(\lambda_2) .$$

Clearly $\Lambda(0) = \log E[1] = 0$, so $\Lambda^*(x) \geq 0x - \Lambda(0) = 0$. Let $x_n \to x$. Then for any $\lambda \in \mathbb{R}$,

$$\liminf_{x_n \to x} \Lambda^*(x_n) \geq \liminf_{x_n \to x} [\lambda x_n - \Lambda(\lambda)] = \lambda x - \Lambda(\lambda)$$

and thus

$$\liminf_{x_n \to x} \Lambda^*(x_n) \geq \sup_{\lambda \in \mathbb{R}} [\lambda x - \Lambda(\lambda)] \triangleq \Lambda^*(x) ,$$

establishing the lower semicontinuity of Λ^*. Thus, Λ^* is a rate function.

The convexity of Λ^* follows by definition as

$$\begin{aligned} \theta\Lambda^*(x_1) + (1-\theta)\Lambda^*(x_2) &= \sup_{\lambda \in \mathbb{R}} \{\theta\lambda x_1 - \theta\Lambda(\lambda)\} + \sup_{\lambda \in \mathbb{R}} \{(1-\theta)\lambda x_2 - (1-\theta)\Lambda(\lambda)\} \\ &\geq \sup_{\lambda \in \mathbb{R}} \{(\theta x_1 + (1-\theta)x_2)\lambda - \Lambda(\lambda)\} = \Lambda^*(\theta x_1 + (1-\theta)x_2) . \end{aligned}$$

(b). By Jensen's inequality,

$$\Lambda(\lambda) = \log E[e^{\lambda X_1}] \geq E[\log e^{\lambda X_1}] = \lambda\overline{x} ,$$

for any $\lambda \in \mathbb{R}$ and thus $\Lambda^*(\overline{x}) = 0$ (this should have been expected in view of (2.2.39)).

Suppose now that $x \in [\overline{x}, \infty)$. Then, for any $\lambda < 0$

$$\lambda x - \Lambda(\lambda) \leq \lambda\overline{x} - \Lambda(\lambda) \leq \Lambda^*(\overline{x}) = 0 ,$$

and (2.2.42) follows since $\Lambda^*(x)$ is non-negative. Moreover, $\Lambda^*(x)$ is nondecreasing for $x > \overline{x}$ since for any $\lambda \geq 0$ the function $g(x) = \lambda x - \Lambda(\lambda)$ is nondecreasing. The proof of (2.2.43) for $x \in (-\infty, \overline{x}]$ is similar.

(c). The identity (2.2.44) follows by interchanging the order of differentiation and integration. This is justified by the dominated convergence theorem since for ϵ small enough

$$E[|X_1|e^{\eta X_1}] \leq \frac{1}{|\epsilon|}(M(\eta+\epsilon) + M(\eta-\epsilon)) < \infty .$$

Let $\Lambda'(\eta) = y$ and consider the function $g(\lambda) \triangleq \lambda y - \Lambda(\lambda)$. As $g(\cdot)$ is a concave function and $g'(\eta) = 0$ clearly $g(\eta) = \sup_{\lambda \in \mathbb{R}} g(\lambda)$ and (2.2.45) is established. □

Proof of Lemma 2.2.2:

(a). The sets $\mathcal{D}_\Lambda$ and $\mathcal{D}_{\Lambda^*}$ are convex since the functions Λ and Λ^* are convex. Further, $\mathcal{D}_\Lambda$ and $\mathcal{D}_{\Lambda^*}$ are intervals since any convex subset of $\mathbb{R}^1$ is an interval. It was already noted that $\Lambda(0) = 0$, so $0 \in \mathcal{D}_\Lambda$. Moreover, $\Lambda'(\cdot)$ is nondecreasing since $\Lambda(\cdot)$ is convex. Now, since $\mathcal{D}_\Lambda^o$ is an interval so is $\mathcal{F}$. Finally, $\mathcal{F} \subseteq \mathcal{D}_{\Lambda^*}$ by (2.2.45).

(b). Since $0 \in \mathcal{D}_\Lambda^o$ certainly $\overline{x} = \Lambda'(0) \in \mathcal{F}$ and if $\mathcal{D}_{\Lambda^*} = \{\overline{x}\}$ the proof is completed. Otherwise, $\mathcal{D}_{\Lambda^*}^o$ is a non-empty interval. Consider now some $x > \overline{x}$ and suppose that $x \in \mathcal{D}_{\Lambda^*}^o$. Recall that $\Lambda^*(x) = \sup_{\lambda \in \mathcal{D}_\Lambda^+} g(\lambda)$ where the concave function $g(\lambda) \stackrel{\Delta}{=} \lambda x - \Lambda(\lambda)$ is continuous within the interval $\mathcal{D}_\Lambda^+ \stackrel{\Delta}{=} \mathcal{D}_\Lambda \cap [0, \infty)$ and differentiable in the non-empty interior of $\mathcal{D}_\Lambda^+$. Consequently, $\Lambda^*(x) = \lim_{r \to \infty} g(\lambda_r)$ for some positive sequence $\lambda_r \in \mathcal{D}_\Lambda^o$ such that $g'(\lambda_r) \geq 0$. Further, the sequence $\{\lambda_r\}$ is bounded since $\Lambda^*(x + \epsilon) < \infty$ for some $\epsilon > 0$, and as such has a limit point, say λ^*. Passing to the convergent subsequence, $\lambda^* \notin \mathcal{D}_\Lambda^o$ implies by the steepness of $\Lambda(\cdot)$ that $\lim_{r \to \infty} g'(\lambda_r) = -\infty$. This contradicts the above requirement that $g'(\lambda_r) \geq 0$. Therefore, $\lambda^* \in \mathcal{D}_\Lambda^o$ implying $x = \Lambda'(\lambda^*)$, i.e., $x \in \mathcal{F}$. A similar proof applies for $x < \overline{x}$ so $\mathcal{D}_{\Lambda^*}^o \subseteq \mathcal{F}$. Since $\mathcal{D}_{\Lambda^*}$ is an interval of non-empty interior it then follows that $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$.

(c). If $0 \in \mathcal{D}_\Lambda^o$ then there exist $\lambda_+ > 0$ and $\lambda_- < 0$ which are both in $\mathcal{D}_\Lambda$. Since for any $\lambda \in \mathbb{R}^1$

$$\frac{\Lambda^*(x)}{|x|} \geq \lambda \mathrm{sign}(x) - \frac{\Lambda(\lambda)}{|x|} ,$$

it follows that

$$\liminf_{|x| \to \infty} \frac{\Lambda^*(x)}{|x|} \geq \min\{\lambda_+, -\lambda_-\} > 0 .$$

Thus, in particular $\Lambda^*(x) \underset{|x| \to \infty}{\longrightarrow} \infty$ and its level sets are bounded. Recall that closed and bounded subsets of $\mathbb{R}^1$ are compact, so Λ^* is indeed a good rate function. □

Exercises:

2.2.1 Suppose that $\mathcal{D}_\Lambda = \mathbb{R}^1$. Prove that $\mu(\{x\}) = e^{-\Lambda^*(x)}$ when $x \in \mathcal{D}_{\Lambda^*} \setminus \mathcal{F}$.

Hint: Show first that there exists $\{\lambda_\ell\}$ such that $|\lambda_\ell| \to \infty$ and $\int_{\mathbb{R}} e^{\lambda_\ell (z-x)} d\mu(z) = e^{-\Lambda^*(x)}$.

2.2.2 Explain why the above proof of (2.2.51) may not work when $x \in \mathcal{D}_{\Lambda^*} \setminus \mathcal{F}$.

Hint: Try distributions with density of the form $Ce^{-\eta|x|}/(1 + |x|^p)$ for appropriate p.

2.2.3 (a). Prove that if $\Lambda(\lambda) = \infty$ for all $\lambda > 0$ and $\overline{x} < \infty$, then $\Lambda^*(x) = 0$ for any $x > \overline{x}$, while $\overline{\mathcal{F}} \subseteq (-\infty, \overline{x}]$.

(b). Suppose now that $\mathcal{D}_\Lambda = (-\infty, \lambda_0]$ where $\lambda_0 > 0$ and $\lim_{\lambda \uparrow \lambda_0} \Lambda'(\lambda) = x_0 < \infty$. Prove that $\Lambda^*(x) = \Lambda^*(x_0) + \lambda_0(x - x_0)$ for any $x > x_0$ while $\overline{\mathcal{F}} \subseteq (-\infty, x_0]$.

2.2.4 Prove that $\Lambda(\cdot)$ is lower semicontinuous.

Hint: Suppose that $M(\lambda) = \infty$ for some $\lambda > 0$ while $\lambda - \delta \in \mathcal{D}_\Lambda^o$ for all $\delta > 0$. Let $dG(x) \triangleq e^{\lambda x} d\mu(x)$ and observe that $\lim_{\delta \to 0} G(\frac{1}{\delta}) = \infty$. By integration by parts show that $M(\lambda - \delta) \geq e^{-1} G(\frac{1}{\delta}) \to \infty$. Conclude the proof using the convexity of $\Lambda(\cdot)$.

2.2.5 (a). Prove that $\Lambda(\lambda)$ is C^∞ in $\mathcal{D}_\Lambda^o$ and that $\Lambda^*(x)$ is strictly convex and C^∞ in $\mathcal{F}^o$.

Hint: Show that $x = \Lambda'(\eta) \in \mathcal{F}^o$ implies that $\Lambda''(\eta) > 0$.

(b). Construct an example where $\mathcal{D}_\Lambda = \mathbb{R}^1$ while $\Lambda^*(\cdot)$ is discontinuous.

Hint: Use a binary valued random variable.

2.2.6 Show that

(a) For X_1 a Poisson(θ) random variable $\Lambda^*(x) = \theta - x + x \log(\frac{x}{\theta})$ for $x \geq 0$ and $\Lambda^*(x) = \infty$ otherwise.

(b) For X_1 a Bernoulli(p) random variable $\Lambda^*(x) = x \log(\frac{x}{p}) + (1 - x) \log(\frac{1-x}{1-p})$ for $x \in [0, 1]$ and $\Lambda^*(x) = \infty$ otherwise.

(c) For X_1 an Exponential (θ) random variable $\Lambda^*(x) = \theta x - 1 - \log(\theta x)$ for $x > 0$ and $\Lambda^*(x) = \infty$ otherwise.

(d) For X_1 a Normal $(0, \sigma^2)$ random variable $\Lambda^*(x) = \frac{x^2}{2\sigma^2}$.

Verify that a large deviations principle holds in all these cases.

2.2.7 Let (x_-, x_+) be the largest interval around $\overline{x}$ such that $(x_-, x_+) \cap \overline{A} = \emptyset$. Define $I_A \triangleq \min\{\Lambda^*(x_-), \Lambda^*(x_+)\}$ and x^* any point at which this minimum is obtained.

(a). Prove that

$$\lim_{n \to \infty} \frac{1}{n} \log \mu_n(A) = -I_A \tag{2.2.54}$$

whenever $I_A = \infty$.

(b). Prove that (2.2.54) also holds for $I_A < \infty$ if $x^* \in \overline{A^o \cap \mathcal{F}}$.

Hint: Let $x_r \in A^o \cap \mathcal{F}$ be a sequence which converges to x^* and apply (2.2.51) in order to verify that $\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(A) \geq -\lim_{r\to\infty} \Lambda^*(x_r) = -\Lambda^*(x^*)$.

2.2.8 Assume that $a \leq X_1 \leq b$

(a) Show that for any λ

$$M(\lambda) \leq \frac{b-\overline{x}}{b-a} e^{\lambda a} + \frac{\overline{x}-a}{b-a} e^{\lambda b} .$$

This is the main ingredient of Hoeffding's inequality.

(b) Use (a) to show that for any $a \leq x \leq b$

$$\Lambda^*(x) \geq H(\frac{x-a}{b-a} | \frac{\overline{x}-a}{b-a}) , \tag{2.2.55}$$

where $H(p|p_0) \stackrel{\Delta}{=} p \log(p/p_0) + (1-p)\log((1-p)/(1-p_0))$.

(c) Prove that the inequality (2.2.55) is sharp.

2.2.9 Assume that $\overline{x} = E[X_1]$ is finite, $X_1 \leq b$ and $Var(X_1) \leq \sigma^2$

(a) Show that for any $\lambda > 0$

$$M(\lambda) \leq e^{\lambda \overline{x}} \left\{ \frac{(b-\overline{x})^2}{(b-\overline{x})^2 + \sigma^2} e^{-\lambda\sigma^2/(b-\overline{x})} + \frac{\sigma^2}{(b-\overline{x})^2 + \sigma^2} e^{\lambda(b-\overline{x})} \right\} .$$

This is the main ingredient of Bennet's inequality.

(b) Use (a) to show that

$$\Lambda^*(x) \geq H(p_x | p_{\overline{x}}) , \tag{2.2.56}$$

for any $\overline{x} \leq x \leq b$ where $p_x \stackrel{\Delta}{=} \frac{(b-\overline{x})(x-\overline{x})+\sigma^2}{(b-\overline{x})^2+\sigma^2}$.

(c) Prove that the inequality (2.2.56) is sharp.

2.3 A general large deviations principle in $\mathbb{R}^d$

Cramer's theorem 2.2.1 possesses a multivariate counterpart dealing with the large deviations of the empirical means of i.i.d real vectors in $\mathbb{R}^d$. Actually, the essential elements in the proof of this theorem extend to a more general class of dependent random vectors. This is explored here, where emphasis is put on the new points in which $\mathbb{R}^d$ differs from $\mathbb{R}^1$ with an eye towards possible infinite dimensional extensions.

Key applications are of-course Cramer's theorem in $\mathbb{R}^d$ which is presented in Theorem 2.3.2 and its consequence – Sanov's Theorem for finite alphabets (see Corollary 2.3.1 and exercise 2.3.1). Some simple non i.i.d. applications are left as exercises 2.3.6, 2.3.8 and 2.3.9 while Section 2.4 is devoted to another class of key applications – the large deviations of Markov chains over finite alphabets.

The set-up considered here consists of a sequence of random vectors $Z_n \in \mathbb{R}^d$ with laws μ_n and logarithmic moment generating functions

$$\Lambda_n(\lambda) \stackrel{\Delta}{=} \log E\left[e^{<\lambda, Z_n>}\right] \tag{2.3.57}$$

where $<\lambda, x> \stackrel{\Delta}{=} \sum_{j=1}^d \lambda^j x^j$ is the usual scalar product in $\mathbb{R}^d$ with x^j the j-th coordinate of the point $x \in \mathbb{R}^d$ and $|x| \stackrel{\Delta}{=} \sqrt{<x, x>}$ the usual Euclidean norm.

The existence of the limit of the properly scaled logarithmic moment generating functions indicates that μ_n may satisfy a large deviations principle. Specifically, the following assumption prevails throughout this section.

Assumption 2.3.1 *There exists a sequence of constants $a_n \underset{n\to\infty}{\longrightarrow} 0$ such that for each $\lambda \in \mathbb{R}^d$, the limit*

$$\Lambda(\lambda) \stackrel{\Delta}{=} \lim_{n\to\infty} a_n \Lambda_n(a_n^{-1}\lambda) \tag{2.3.58}$$

exists (possibly as a point in $[-\infty, \infty]$). Further, $0 \in \mathcal{D}_\Lambda^o$ where $\mathcal{D}_\Lambda$ is the domain in which $\Lambda(\cdot) < \infty$.

For example, when μ_n is the law governing the empirical mean $\hat{S}_n$ of the i.i.d. random vectors $X_i \in \mathbb{R}^d$ then for any $n \in Z^+$

$$\frac{1}{n}\Lambda_n(n\lambda) = \Lambda(\lambda) \stackrel{\Delta}{=} \log E[e^{<\lambda, X_1>}] \tag{2.3.59}$$

and the above assumption holds whenever $0 \in \mathcal{D}_\Lambda^o$.

Definition 2.3.1 *The Fenchel-Legendre transform of* $\Lambda(\lambda)$ *is*

$$\Lambda^*(x) \stackrel{\Delta}{=} \sup_{\lambda \in \mathbb{R}^d} (< \lambda, x > -\Lambda(\lambda)), \tag{2.3.60}$$

with $\mathcal{D}_{\Lambda^*}$ *denoting the domain in which* $\Lambda^*(\cdot) < \infty$.

Cramer's theorem suggests that Λ^* is the natural candidate rate function for governing the large deviations principle associated with μ_n. This is indeed proved in Theorem 2.3.1 under an additional condition. The properties of the functions Λ and Λ^* which are needed for that purpose are summarized in the following lemma whose proof is deferred to the end of this section.

Lemma 2.3.1 *Assume 2.3.1.*

(a). $\Lambda(\lambda)$ *is a convex function,* $\Lambda(\lambda) > -\infty$ *everywhere and* $\Lambda^*(x)$ *is a good, convex rate function.*

(b). $\mathcal{D}_\Lambda$ *and* $\mathcal{D}_{\Lambda^*}$ *are convex sets.*

(c). Suppose that $y = \nabla\Lambda(\eta)$ *for some* $\eta \in \mathcal{D}_\Lambda^o$. *Then*

$$\Lambda^*(y) = < \eta, y > -\Lambda(\eta)\,. \tag{2.3.61}$$

(d). Let y, η *be as in (c) above. Then, for any* $x \neq y$

$$\Lambda_\eta^*(x) > \Lambda_\eta^*(y) = 0\ , \tag{2.3.62}$$

where $\Lambda_\eta^*(\cdot)$ *is the Fenchel-Legendre transform of*

$$\Lambda_\eta(\theta) \stackrel{\Delta}{=} \Lambda(\theta + \eta) - \Lambda(\eta)\ . \tag{2.3.63}$$

Remark: These convex analysis considerations are addressed again in Section ?? in a more abstract setup.

The general large deviations principle in $\mathbb{R}^d$ is now summarized as follows.

Theorem 2.3.1 (Gartner) *Assume 2.3.1. Then*

(a).For any closed set F

$$\limsup_{n \to \infty}\ a_n \log \mu_n(F) \leq -\inf_{x \in F} \Lambda^*(x)\,. \tag{2.3.64}$$

(b). For any open set G

$$\liminf_{n\to\infty} a_n \log \mu_n(G) \geq - \inf_{x\in G\cap\overline{\mathcal{F}}} \Lambda^*(x)\,, \tag{2.3.65}$$

where $\mathcal{F} \stackrel{\Delta}{=} \{x : x = \nabla\Lambda(\lambda) \ \ \textit{for some} \ \ \lambda \in \mathcal{D}_\Lambda^o\}$.

(c). If $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$ *then the family of probability measures* μ_n *satisfies a large deviations principle controlled by the good rate function* $\Lambda^*(x)$.

The following lemma, whose proof is also deferred to the end of this section, makes the above theorem applicable by stating explicit conditions on Λ under which $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$.

Lemma 2.3.2 (Ellis) *Suppose that* $\Lambda(\lambda)$ *which satisfies 2.3.1 is a lower semicontinuous function which is differentiable in* $\mathcal{D}_\Lambda^o$ *and moreover it is a* steep *function, namely for any convergent sequence* $\lambda_n \in \mathcal{D}_\Lambda^o$ *whose limit does not belong to* $\mathcal{D}_\Lambda^o$, $\lim_{n\to\infty} |\nabla\Lambda(\lambda_n)| = \infty$. *Then,* $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$.

The proof of Theorem 2.3.1 is given in the sequel, preceded by the statement and derivation of a key application – Cramer's theorem about the empirical means of i.i.d. random vectors in $\mathbb{R}^d$.

Theorem 2.3.2 *Let* μ_n *be the laws governing the empirical means* $\hat{S}_n \stackrel{\Delta}{=} \frac{1}{n}\sum_{i=1}^n X_i$ *where* $X_i \in \mathbb{R}^d$ *are i.i.d. distributed according to the law* μ. *Suppose that the logarithmic moment generating function*

$$\Lambda(\lambda) \stackrel{\Delta}{=} \log E_\mu \left[e^{<\lambda, X_1>}\right] \,, \tag{2.3.66}$$

is a steep lower semicontinuous function which is finite in some open ball centered at the origin (in particular these conditions hold when $\mathcal{D}_\Lambda = \mathbb{R}^d$*). Then, for any measurable set* $\Gamma \subset \mathbb{R}^d$

$$-\inf_{x\in\Gamma^o} \Lambda^*(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(\Gamma) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(\Gamma) \leq -\inf_{x\in\overline{\Gamma}} \Lambda^*(x)\,, \tag{2.3.67}$$

with Λ^* *being the Fenchel-Legendre transform of* Λ.

Proof: Recall that in this case the basic assumption 2.3.1 holds (see (2.3.59)) and indeed Λ of (2.3.58) is given by (2.3.66). In the statement of the theorem it is further assumed that Λ is a steep function. It follows from the definition (2.3.66) that Λ is differentiable in $\mathcal{D}_\Lambda^o$ (by an argument which is similar to the proof of (2.2.44) in Section 2.2). Thus, Lemma 2.3.2 applies and (2.3.67) follows by Theorem 2.3.1. □

Remark The assumption of lower semi-continuity may be relaxed. For details, c.f. exercise 2.3.3.

The essential ingredients for the proof of Theorem 2.3.1 are those presented in Section 2.2 in the course of proving Theorem 2.2.1, namely – the exponential form of Markov's inequality is applied for deriving the upper bound and an exponential change of measure is used for deriving the lower bound. However, here one encounters two new obstacles which slightly complicate both parts of the proof.

Proof of Theorem 2.3.1:

(a). In $\mathbb{R}^d$ the monotonicity of Λ^* stated in Lemma 2.2.1 part (b) is somewhat lost. Thus, the strategy of containing $\overline{A}$ by two half-spaces A_- and A_+ is not as useful as it is in $\mathbb{R}^1$. Instead, here one uses Markov's inequality to obtain tight upper bounds for all small closed balls. Then, compact sets are covered by an appropriate finite collection of small enough balls and the upper bound follows for compact sets by the union of events bound.

As mentioned in Section 1.1, proving (2.3.64) is equivalent to proving that for any $\delta > 0$ and any closed set $F \subset \mathbb{R}^d$

$$\limsup_{n\to\infty} a_n \log \mu_n(F) \leq \delta - \inf_{x\in F} I^\delta(x) \tag{2.3.68}$$

where

$$I^\delta(x) \triangleq \begin{cases} \Lambda^*(x) - \delta & x \in \mathcal{D}_{\Lambda^*} \\ \frac{1}{\delta} & x \notin \mathcal{D}_{\Lambda^*} \end{cases}$$

Fix now $\delta > 0$ and an arbitrary **compact** set Γ. For any $q \in \Gamma$ choose $\lambda_q \in \mathcal{D}_\Lambda$ for which

$$< \lambda_q, q > - \Lambda(\lambda_q) \geq I^\delta(q) \tag{2.3.69}$$

This is always possible by the definitions of Λ^* and I^δ. Choose now $\rho_q > 0$ such that $\rho_q|\lambda_q| \leq \delta$ and let B_{q,ρ_q} be the open ball with center at the point q and radius ρ_q with $\overline{B}_{q,\rho_q}$ the corresponding closed ball.

Then, for any n and any $q \in \Gamma$

$$\mu_n(B_{q,\rho_q}) \leq \exp\left(- \inf_{x\in\overline{B}_{q,\rho_q}} \{a_n^{-1} < \lambda_q, x >\} \right) E\left[\exp(a_n^{-1} < \lambda_q, Z_n >)\right] .$$

Thus,

$$a_n \log \mu_n(B_{q,\rho_q}) \leq - \inf_{x\in\overline{B}_{q,\rho_q}} < \lambda_q, x > + a_n \Lambda_n(a_n^{-1}\lambda_q) \leq \delta - < \lambda_q, q > + a_n \Lambda_n(a_n^{-1}\lambda_q) . \tag{2.3.70}$$

As Γ is a compact set, one can extract from the open cover $\bigcup_{q\in\Gamma} B_{q,\rho_q}$ of Γ a finite cover which consists of $N < \infty$ (depending only on Γ and δ) such balls with centers $q_1, \ldots, q_N$ in Γ. By the union of events bound and (2.3.70),

$$a_n \log \mu_n(\Gamma) \leq a_n \log N + \delta - \min_{i=1,\ldots,N} \{< \lambda_{q_i}, q_i > - a_n \Lambda_n(a_n^{-1}\lambda_{q_i})\} .$$

Since $a_n \to 0$ as $n \to \infty$ while $a_n \Lambda_n(a_n^{-1}\lambda_{q_i}) \to \Lambda(\lambda_{q_i})$ (uniformly over $i = 1, \ldots, N$), one obtains

$$\limsup_{n\to\infty} a_n \log \mu_n(\Gamma) \leq \delta - \min_{i=1,\ldots,N} \{< \lambda_{q_i}, q_i > - \Lambda(\lambda_{q_i})\} \leq \delta - \min_{i=1,\ldots,N} I^{\delta}(q_i) ,$$

where the last inequality follows from (2.3.69). As $q_i \in \Gamma$ the upper bound (2.3.68) is thus established for all compact sets.

This upper bound is extended to all closed sets in $\mathbb{R}^d$ by showing that μ_n is an *exponentially tight* family of probability measures and using Lemma 1.1.1. Specifically, it is shown in the sequel that for any $\alpha < \infty$ there exists ρ_α large enough such that for the compact set $K_\alpha \triangleq \overline{B}_{0,\rho_\alpha}$

$$\limsup_{n\to\infty} a_n \log \mu_n(K_\alpha^c) < -\alpha \tag{2.3.71}$$

For that purpose, observe first that $\overline{B}^c_{0,\rho d} \subset \cup_{j=1}^d \{x : |x^j| \geq \rho\}$. Therefore, by the union of events bound

$$\begin{aligned} \mu_n(\overline{B}^c_{0,\rho d}) &\leq \mu_n(\cup_{j=1}^d \{x : |x^j| \geq \rho\}) \\ &\leq \sum_{j=1}^d \left\{\mu_n^j([\rho,\infty)) + \mu_n^j((-\infty,-\rho])\right\} \end{aligned} \tag{2.3.72}$$

where μ_n^j are the laws of Z_n^j, the coordinates of the random vector Z_n. As $\mathcal{D}_\Lambda$ contains an open ball around the origin there exist $\theta_j^+ > 0$ and $\theta_j^- < 0$ such that $\Lambda(\theta_j^+ \mathbf{u}_j) < \infty$ and $\Lambda(\theta_j^- \mathbf{u}_j) < \infty$ where $\mathbf{u_j}$ denotes the j-th unit vector in $\mathbb{R}^d$ for $j = 1, \ldots, d$. By the exponential form of Markov's inequality, for $j = 1, \ldots, d$

$$\limsup_{n\to\infty} a_n \log \mu_n^j([\rho,\infty)) \leq -\theta_j^+ \rho + \limsup_{n\to\infty} a_n \Lambda_n(a_n^{-1}\theta_j^+ \mathbf{u}_j) = -\theta_j^+ \rho + \Lambda(\theta_j^+ \mathbf{u}_j) . \tag{2.3.73}$$

Similarly,

$$\limsup_{n\to\infty} a_n \log \mu_n^j((-\infty,-\rho]) \leq \theta_j^- \rho + \Lambda(\theta_j^- \mathbf{u}_j) . \tag{2.3.74}$$

Now, the inequality (2.3.71) results by combining (2.3.72), (2.3.73) and (2.3.74) and considering $\rho \to \infty$.

(b). Focusing now on establishing the lower bound (2.3.65) for any open set, it suffices to prove

$$\lim_{\delta \to 0} \liminf_{n \to \infty} a_n \log \mu_n(B_{y,\delta}) \geq -\Lambda^*(y) , \tag{2.3.75}$$

for any $y \in \mathcal{F}$. Indeed, (2.3.75) then holds for any $y \in \overline{\mathcal{F}}$ and (2.3.65) follows by the same argument encountered in that leads from (2.2.47) to (2.2.51) in the proof of Theorem 2.2.1.

Fix now $y = \nabla\Lambda(\eta) \in \mathcal{F}$ with $\eta \in \mathcal{D}_\Lambda^o$. Then, for all n large enough, $\Lambda_n(a_n^{-1}\eta) < \infty$ and the "associated" probability measures

$$d\tilde{\mu}_n(z) = \exp\left[a_n^{-1} < \eta, z > -\Lambda_n(a_n^{-1}\eta)\right] d\mu_n(z) \tag{2.3.76}$$

are well defined. Clearly,

$$\begin{aligned} a_n \log \mu_n(B_{y,\delta}) &= a_n\Lambda_n(a_n^{-1}\eta) - < \eta, y > + a_n \log \int_{z \in B_{y,\delta}} \exp(a_n^{-1} < \eta, y - z >) \, d\tilde{\mu}_n(z) \\ &\geq a_n\Lambda_n(a_n^{-1}\eta) - < \eta, y > - |\eta|\delta + a_n \log \tilde{\mu}_n(B_{y,\delta}) \end{aligned} \tag{2.3.77}$$

Therefore,

$$\begin{aligned} \lim_{\delta \to 0} \liminf_{n \to \infty} a_n \log \mu_n(B_{y,\delta}) &\geq \Lambda(\eta) - < \eta, y > + \lim_{\delta \to 0} \liminf_{n \to \infty} a_n \log \tilde{\mu}_n(B_{y,\delta}) \\ &= -\Lambda^*(y) + \lim_{\delta \to 0} \liminf_{n \to \infty} a_n \log \tilde{\mu}_n(B_{y,\delta}) \end{aligned} \tag{2.3.78}$$

where the above equality follows by (2.3.61).

Here, a new obstacle stems from the removal of the independence assumption. Indeed, $E_{\tilde{\mu}_n}(Z_n) \to y$ as $n \to \infty$, but now one still has to establish the appropriate analog of the weak law of large numbers. This is handled in the sequel by applying the *large deviations upper bound* for the "associated" family of measures $\tilde{\mu}_n$. Indeed, the proof of (2.3.75) is completed by showing that for any $\delta > 0$

$$\limsup_{n \to \infty} a_n \log \tilde{\mu}_n(B_{y,\delta}^c) < 0 . \tag{2.3.79}$$

For that purpose let $\Lambda_{n,\eta}(\cdot)$ denotes the logarithmic moment generating function corresponding to the law $\tilde{\mu}_n$. Then, for any $\theta \in \mathbb{R}^d$

$$a_n\Lambda_{n,\eta}(a_n^{-1}\theta) \triangleq a_n \log \left[\int_{\mathbb{R}^d} e^{a_n^{-1} < \theta, z >} \, d\tilde{\mu}_n(z)\right] = a_n\Lambda_n(a_n^{-1}(\theta + \eta)) - a_n\Lambda_n(a_n^{-1}\eta) \to \Lambda_\eta(\theta)$$

as $a_n\Lambda_n(a_n^{-1}\eta) < \infty$ for all n large (recall that $\eta \in \mathcal{D}_\Lambda$). Moreover, $\Lambda_\eta(0) = 0$ and $\Lambda_\eta(\theta) < \infty$ for all $|\theta|$ small enough since $\eta \in \mathcal{D}_\Lambda^o$ (c.f. (2.3.63)). Thus, a large deviations *upper bound* of the form of (2.3.64) holds for the sequence of measures $\tilde{\mu}_n$. In particular, for the closed set $B_{y,\delta}^c$ it yields

$$\limsup_{n\to\infty} a_n \log \tilde{\mu}_n(B_{y,\delta}^c) \leq -\inf_{x\in B_{y,\delta}^c} \Lambda_\eta^*(x). \tag{2.3.80}$$

Moreover, $\Lambda_\eta^*(x) > 0$ for any $x \neq y$ in view of (2.3.62) and paralleling the proof of part (a) of Lemma 2.3.1 one easily shows that Λ_η^* is a good rate function. Therefore, $\inf_{x\in B_{y,\delta}^c} \Lambda_\eta^*(x) > 0$ for all $\delta > 0$ and (2.3.80) implies (2.3.79), concluding the proof of the lower bound (2.3.65).

(c). The large deviations principle follows by combining (2.3.64) and (2.3.65) as $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$ implies that for any G

$$-\inf_{x\in G\cap\overline{\mathcal{F}}} \Lambda^*(x) \geq -\inf_{x\in G\cap\mathcal{D}_{\Lambda^*}} \Lambda^*(x) = -\inf_{x\in G} \Lambda^*(x)$$

□

Remarks:

(a). The proof above actually extends beyond $\mathbb{R}^d$ and in principle is applicable for any metric space (as shown in Chapter ??). However, two points of caution are that the exponential tightness has to be proved on a case by case basis and that in infinite dimensional spaces Λ is rarely differentiable and therefore the analog of Lemma 2.3.2 typically fails to hold.

(b). In the current set-up the condition $0 \in \mathcal{D}_\Lambda^o$ which is needed for the above proof does not follow as a consequence of $\mathcal{D}_{\Lambda^*} \subseteq \overline{\mathcal{F}}$. For this reason it is incorporated in Assumption 2.3.1. Note however that the condition $0 \in \mathcal{D}_\Lambda^o$ is not required at all for proving the upper bound (2.3.64) for compact sets.

As mentioned in Section 2.1, Sanov's theorem for finite alphabets, Theorem 2.1.1, may be deduced as a consequence of Cramer's theorem 2.3.2. Indeed, note that the empirical mean of the random vectors $Y_i \stackrel{\Delta}{=} [1_{X_i=a_1}, 1_{X_i=a_2} \ldots, 1_{X_i=a_{|\Sigma|}}]$ equals $L_n^{\mathbf{X}}$, the empirical measure of the i.i.d. random variables $X_1, \ldots, X_n$ over the finite alphabet Σ. Moreover, as Y_i are bounded, here $\mathcal{D}_\Lambda = \mathbb{R}^{|\Sigma|}$ and one obtains the following corollary of Theorem 2.3.2.

Corollary 2.3.1 *For any set Γ of probability vectors in $\mathbb{R}^{|\Sigma|}$*

$$-\inf_{\nu\in\Gamma^o} \Lambda^*(\nu) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) \leq -\inf_{\nu\in\overline{\Gamma}} \Lambda^*(\nu) \tag{2.3.81}$$

where Λ^* is the Fenchel-Legendre transform of the logarithmic moment generating function

$$\Lambda(\lambda) = \log E_\mu(e^{<\lambda, Y_1>}) = \log \sum_{i=1}^{|\Sigma|} e^{\lambda_i} \mu(a_i) \tag{2.3.82}$$

and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{|\Sigma|}) \in \mathbb{R}^{|\Sigma|}$.

Remark: Comparing the above corollary to Theorem 2.1.1 it is tempting to conjecture that $\Lambda^*(\cdot) = H(\cdot|\mu)$. Indeed, this is proved in exercise 2.3.1. Actually, as shown in Section ?? the rate function controlling a large deviations principle in $\mathbb{R}^d$ is always unique, thus this result is not surprising.

Proof of Lemma 2.3.1:

(a). Since Λ_n are convex functions (see the proof of part (a) of Lemma 2.2.1) so are $a_n \Lambda_n(a_n^{-1}\cdot)$ and their limit $\Lambda(\cdot)$ is convex as well. Moreover, $\Lambda_n(0) = 0$ and therefore $\Lambda(0) = 0$ implying that Λ^* is non-negative. Both the convexity and the lower semicontinuity of Λ^* follow from its definition via (2.3.60) (see proof of part (a) of Lemma 2.2.1).

Now if $\Lambda(\lambda) = -\infty$ for some $\lambda \in \mathbb{R}^d$ then by convexity $\Lambda(\alpha\lambda) = -\infty$ for all $\alpha \in (0, 1]$. Moreover, since $\Lambda(0) = 0$ it follows by convexity that $\Lambda(-\alpha\lambda) = \infty$ for all $\alpha \in (0, 1]$ contradicting the assumption that $0 \in \mathcal{D}_\Lambda^o$.

Since $0 \in \mathcal{D}_\Lambda^o$, it follows that $\overline{B}_{0,\delta} \subset \mathcal{D}_\Lambda^o$ for some $\delta > 0$ and $C = \sup_{\lambda \in \overline{B}_{0,\delta}} \Lambda(\lambda) < \infty$ since the convex function Λ is continuous in $\mathcal{D}_\Lambda^o$. Therefore,

$$\Lambda^*(x) \geq \sup_{\lambda \in \overline{B}_{0,\delta}} \{< \lambda, x > -\Lambda(\lambda)\} \geq \sup_{\lambda \in \overline{B}_{0,\delta}} < \lambda, x > - \sup_{\lambda \in \overline{B}_{0,\delta}} \Lambda(\lambda) = \delta|x| - C \tag{2.3.83}$$

Thus, the level sets $\{x : \Lambda^*(x) \leq \alpha\}$ are clearly bounded within a closed ball around the origin (of radius $(\alpha + C)/\delta$) and Λ^* is necessarily a good rate function.

(b). The convexity of $\mathcal{D}_\Lambda$ and $\mathcal{D}_{\Lambda^*}$ is merely a consequence of the convexity of Λ and Λ^* respectively.

(c). Clearly $\Lambda^*(y) \geq [< \eta, y > -\Lambda(\eta)]$. Assume that this inequality is strict, i.e., for some $\lambda \in \mathcal{D}_\Lambda$

$$[< \lambda, y > -\Lambda(\lambda)] > [< \eta, y > -\Lambda(\eta)]. \tag{2.3.84}$$

Since $\mathcal{D}_\Lambda$ is a convex set,

$$g(\alpha) \stackrel{\Delta}{=} \alpha < \lambda - \eta, y > -\Lambda(\eta + \alpha(\lambda - \eta)) + < \eta, y > \quad \alpha \in [0, 1] \tag{2.3.85}$$

is a finite valued concave function. Thus, by concavity

$$g(1) - g(0) \leq \liminf_{\alpha \downarrow 0} \frac{g(\alpha) - g(0)}{\alpha} =< \lambda - \eta, y - \nabla\Lambda(\eta) >= 0 \,, \tag{2.3.86}$$

where the last equality follows from the assumption $y = \nabla\Lambda(\eta)$. However, the inequality above contradicts (2.3.84) and therefore (2.3.61) is established.

(d). The function Λ_η^* is non-negative (being the Fenchel-Legendre transform of Λ_η where $\Lambda_\eta(0) = 0$) and moreover

$$\Lambda_\eta^*(x) = \sup_{\theta \in \mathbb{R}^d} \{< \theta, x > -\Lambda(\theta + \eta)\} + \Lambda(\eta) = \Lambda^*(x) - < \eta, x > + \Lambda(\eta) \,. \tag{2.3.87}$$

Thus, $\Lambda_\eta^*(y) = 0$ by (2.3.61) and $\Lambda_\eta^*(x) \geq \Lambda_\eta^*(y)$ for any $x \neq y$. If $\Lambda_\eta^*(x) = 0$ for some x then for any $\theta \in \mathbb{R}^d$

$$< \theta, x > \ \leq \Lambda_\eta(\theta) \ . \tag{2.3.88}$$

In particular also

$$< \theta, x > \leq \lim_{n \to \infty} \frac{\Lambda_\eta^*(\theta/n)}{(1/n)} =< \theta, \nabla\Lambda_\eta(0) > \,, \tag{2.3.89}$$

where $\nabla\Lambda_\eta(0) = \nabla\Lambda(\eta) = y$. Since the inequality (2.3.89) holds for all $\theta \in \mathbb{R}^d$, necessarily $x = \nabla\Lambda_\eta(0) = y$. Thus, $\Lambda_\eta^*(x) > 0$ for any $x \neq y$ and (2.3.62) is established. □

Proof of Lemma 2.3.2:

Since $\mathcal{D}_\Lambda^o$ is not empty, by (2.3.61) so is $\mathcal{D}_{\Lambda^*}$. Further, if $\mathcal{D}_{\Lambda^*} = \{\mathbf{x}\}$ then necessarily $\mathbf{x} = \nabla\Lambda(0) \in \mathcal{F}$ and the proof is then complete. Therefore, we may assume from here on that the convex set $\mathcal{D}_{\Lambda^*}$ contains at least one line so its **relative interior**

$$\text{ri } \mathcal{D}_{\Lambda^*} \stackrel{\Delta}{=} \{x \in \mathcal{D}_{\Lambda^*} \ : \ y \in \mathcal{D}_{\Lambda^*} \ \Rightarrow \ x - \epsilon(y - x) \in \mathcal{D}_{\Lambda^*} \text{ for some } \epsilon > 0\}$$

is non-empty and moreover

$$\mathcal{D}_{\Lambda^*} = \overline{\text{ri } \mathcal{D}_{\Lambda^*}} \,. \tag{2.3.90}$$

The subdifferential $\partial\Lambda(\lambda)$ of the convex function $\Lambda(\cdot)$ at a point $\lambda \in \mathbb{R}^d$, is defined as

$$\partial\Lambda(\lambda) \stackrel{\Delta}{=} \{x : \quad \Lambda(\theta) \geq \Lambda(\lambda) + < x, \theta - \lambda > \quad \forall\theta \in \mathbb{R}^d\} = \{x :< \lambda, x > -\Lambda^*(x) \geq \Lambda(\lambda)\} \,, \tag{2.3.91}$$

where the second equality is a direct consequence of the definition of the Fenchel-Legendre transform. It is easy to check that $\partial\Lambda(\lambda) = \emptyset$ for $\lambda \notin \mathcal{D}_\Lambda$ (recall that $\Lambda > -\infty$ everywhere) while

otherwise $\partial\Lambda(\eta) = \{x : \Lambda^*_\eta(x) = 0\}$ (where Λ^*_η is defined in (2.3.63)). Since Λ is differentiable throughout $\mathcal{D}^o_\Lambda$ (by assumption), it follows by (2.3.61) and (2.3.62) that $\partial\Lambda(\eta) = \{\nabla\Lambda(\eta)\}$ for any $\eta \in \mathcal{D}^o_\Lambda$.

The proof of the lemma is now divided into the following two steps:

(a). Since Λ is lower semicontinuous, for any $x \in \text{ri } \mathcal{D}_{\Lambda^*}$ there exists $\lambda \in \mathbb{R}^d$ such that $x \in \partial\Lambda(\lambda)$.

(b). Since Λ is a steep function, $\partial\Lambda(\lambda) = \emptyset$ for any $\lambda \in \mathcal{D}_\Lambda \setminus \mathcal{D}^o_\Lambda$.

When combined, these two claims result with

$$\text{ri } \mathcal{D}_{\Lambda^*} \subseteq \bigcup_{\lambda\in\mathbb{R}^d} \partial\Lambda(\lambda) = \bigcup_{\lambda\in\mathcal{D}^o_\Lambda} \{\nabla\Lambda(\lambda)\} = \mathcal{F}\,, \tag{2.3.92}$$

which, together with (2.3.90), amount to the proof of the lemma.

(a). Fix a point $x \in \text{ri } \mathcal{D}_{\Lambda^*}$ and define the function

$$g(y) \stackrel{\Delta}{=} \inf_{\delta>0} \frac{\Lambda^*(x+\delta y) - \Lambda^*(x)}{\delta} = \lim_{\delta\downarrow 0} \frac{\Lambda^*(x+\delta y) - \Lambda^*(x)}{\delta}\,, \tag{2.3.93}$$

where the convexity of Λ^* results with a monotonicity in δ which in turn implies the above equality (and that the above limit exists). For the same reason $g(y)$ is a convex function and the set $\mathcal{E}_g \stackrel{\Delta}{=} \{(y,\xi) : \xi \geq g(y)\} \subseteq \mathbb{R}^d \times \mathbb{R}$ is a convex set. Further, $g(\alpha y) = \alpha g(y)$ for all $\alpha \geq 0$ and in particular $g(0) = 0$. Observe that $g(y) = \infty$ when $x + \delta y \notin \mathcal{D}_{\Lambda^*}$ for all $\delta > 0$. Consider therefore those directions y such that $g(y) < \infty$. Since $x \in \text{ri } \mathcal{D}_{\Lambda^*}$, it then follows that for some $\epsilon > 0$ the whole line segment $x + \beta y$ for $|\beta| \leq \epsilon$ is in $\mathcal{D}_{\Lambda^*}$. Let $\hat{y} \stackrel{\Delta}{=} \epsilon y$ so by the convexity of Λ^*

$$\Lambda^*(x) < (1-\delta)\Lambda^*(x+\delta\hat{y}) + \delta\Lambda^*(x-(1-\delta)\hat{y}) < \infty \tag{2.3.94}$$

for all $\delta \in [0,1]$. This implies

$$g(\hat{y}) \geq \lim_{\delta\downarrow 0}[\Lambda^*(x+\delta\hat{y}) - \Lambda^*(x-(1-\delta)\hat{y})] \geq -\Lambda^*(x-\hat{y}) = -\Lambda^*(x-\epsilon y) > -\infty \tag{2.3.95}$$

where the last two inequalities follow by the non-negativity and upper semicontinuity of Λ^*, and the fact that $x - \epsilon y \in \mathcal{D}_{\Lambda^*}$. Since $g(y) = \frac{g(\hat{y})}{\epsilon}$ it follows from the above that $g(y) > -\infty$ for all y. Since $g(y) = |y|g(y/|y|)$ and $\inf_{\{z:|z|=1\}} g(z) > -\infty$ it now follows that $\liminf_{y\to 0} g(y) \geq 0$, implying that $(0,-1) \notin \overline{\mathcal{E}}_g$. The set $\overline{\mathcal{E}}_g$ is closed, convex and non-empty (for example $(0,1) \in \overline{\mathcal{E}}_g$). Thus, there

exists a hyperplane in $\mathbb{R}^d \times \mathbb{R}$ which strictly separates the point $(0,-1)$ and the set $\mathcal{E}_g$ (this is a particular instance of the Hahn-Banach theorem quoted in Appendix ??). Specifically, there exist $\lambda \in \mathbb{R}^d$ and $\rho \in \mathbb{R}$ such that

$$< \lambda, 0 > + \rho \; = \; \rho \; > < \lambda, y > - \xi\rho \quad \forall (y,\xi) \in \mathcal{E}_g \; . \tag{2.3.96}$$

Considering $y = 0$ it is clear that $\rho > 0$ and then by specializing (2.3.96) to $\xi = g(y)$ one obtains that $g^*(\frac{\lambda}{\rho}) \leq 1$ where g^* is the Fenchel-Legendre transform of $g(y)$. Observe now that g^* assumes only the values 0 or ∞ as

$$g^*(\lambda) \stackrel{\Delta}{=} \sup_y (< \lambda, y > - g(y)) = \sup_{\alpha \geq 0} \alpha [\sup_{\{z:|z|=1\}} (< \lambda, z > - g(z)) \;] = \begin{cases} 0 & \lambda \in C \\ \infty & \lambda \notin C \end{cases} \tag{2.3.97}$$

where $C = \{\lambda : |z| = 1 \; \Rightarrow \; g(z) \geq < \lambda, z >\}$. Thus, the set C must be non-empty or equivalently there exists $\lambda_0 \in \mathbb{R}^d$ such that $g(y) \geq < \lambda_0, y >$ for all $y \in \mathbb{R}^d$.

Considering now (2.3.93) one obtains

$$\Lambda^*(z) - \Lambda^*(x) \geq < \lambda_0, z - x >$$

for all $z \in \mathbb{R}^d$. Therefore, also

$$< \lambda_0, x > - \Lambda^*(x) = \sup_{z \in \mathbb{R}^d} (< \lambda_0, z > - \Lambda^*(z)) \; \geq \Lambda(\lambda_0) \; , \tag{2.3.98}$$

where the above inequality is a property of the Legendre transform of any convex, lower semicontinuous function f such that $f(\cdot) > -\infty$ everywhere and $f(\cdot) \not\equiv \infty$ (consider Section ?? for a proof). Indeed, it is assumed in the statement of this lemma that $\Lambda(\cdot)$ is lower semicontinuous and the other conditions mentioned above are satisfied in view of part (a) of Lemma 2.3.1. The inequality (2.3.98) amounts to $x \in \partial\Lambda(\lambda_0)$ for some $\lambda_0 \in \mathcal{D}_\Lambda$.

(b). Suppose there exists a point $\eta \in \mathcal{D}_\Lambda \setminus \mathcal{D}_\Lambda^o$ such that $\partial\Lambda(\eta)$ is non-empty. Then, $\partial\Lambda(\eta) = \{x : \Lambda_\eta^*(x) = 0\}$ is a non-empty closed convex set (since Λ_η^* is a convex rate function). Suppose that an infinite line, say $\{z_0 + tz\}_{t \in \mathbb{R}}$ is contained in $\partial\Lambda(\eta)$. Then, for all $\theta \in \mathbb{R}^d$ and any $t \in \mathbb{R}$,

$$\Lambda(\theta) \geq \Lambda(\eta) + < \theta - \eta, z_0 > + t < \theta - \eta, z > \tag{2.3.99}$$

which is possible only if $\mathcal{D}_\Lambda \subseteq \{\theta \; : < \theta, z >= 0\}$ contradicting Assumption 2.3.1. From the above, it follows that $\partial\Lambda(\eta)$ is a convex closed set which does not contain infinite lines and as such by [24],

18.5.3, it contains an exposed point, i.e. there exists $x \in \partial\Lambda(\eta)$ and a vector $\mathbf{v} \in \mathbb{R}^d$ such that

$$< \mathbf{v}, x > > < \mathbf{v}, z > \qquad \forall z \in \partial\Lambda(\eta), \ z \neq x . \tag{2.3.100}$$

Let the normal cone to $\mathcal{D}_\Lambda$ at η be defined as

$$\mathcal{N} \stackrel{\Delta}{=} \{\mathbf{n} : < \lambda - \eta, \mathbf{n} > \leq 0 \quad \text{for all} \ \ \lambda \in \mathcal{D}_\Lambda\} \tag{2.3.101}$$

and note that $\mathcal{N}$ is non-empty since $\mathcal{D}_\Lambda$ is a convex set of non-empty interior and $\eta \in \mathcal{D}_\Lambda \setminus \mathcal{D}_\Lambda^o$. Then, for any $\mathbf{n} \in \mathcal{N}$ and any $\lambda \in \mathbb{R}^d$

$$\Lambda(\lambda) \geq \Lambda(\eta) + < \lambda - \eta, \quad x > \geq \Lambda(\eta) + < \lambda - \eta, \quad x + \mathbf{n} > ,$$

so that $x + \mathbf{n} \in \partial\Lambda(\eta)$ where $x \in \partial\Lambda(\eta)$ is the exposed point defined above. Therefore, in particular, by (2.3.100)

$$< \mathbf{v} , \mathbf{n} > < \ 0 \quad \forall \mathbf{n} \in \mathcal{N} . \tag{2.3.102}$$

We next claim that $\eta + \delta\mathbf{v} \in \mathcal{D}_\Lambda^o$ for all $\delta > 0$ small enough. Indeed, assume otherwise, then by [24], 23.7.1 there exists $\mathbf{n} \in \mathcal{N}$ such that

$$\sup_{\lambda \in \mathcal{D}_\Lambda} < \lambda, \mathbf{n} > \leq < \eta, \mathbf{n} > \leq < \eta + \delta\mathbf{v}, \mathbf{n} > \tag{2.3.103}$$

contradicting the inequality (2.3.102) above.

Choose now a sequence $\lambda_n = \eta + \delta_n \mathbf{v} \in \mathcal{D}_\Lambda^o$ such that $\delta_n \to 0$. Since $\nabla\Lambda(\lambda_n) \in \partial\Lambda(\lambda_n)$ for any n it follows that

$$\Lambda(\theta) \geq \Lambda(\lambda_n) + < \theta - \lambda_n, \ \nabla\Lambda(\lambda_n) > , \qquad \forall \theta \in \mathbb{R}^d \tag{2.3.104}$$

Because of the assumption that Λ is a steep function, $\epsilon_n \stackrel{\Delta}{=} 1/|\nabla\Lambda(\lambda_n)| \to 0$ as $n \to \infty$ and $\epsilon_n \nabla\Lambda(\lambda_n)$ has a limit point $y \in \mathbb{R}^d$ with $|y| = 1$. Passing to the convergent subsequence $\{\lambda_n\}$, for any n large enough and any $\theta \in \mathbb{R}^d$,

$$\begin{aligned} \Lambda(\theta) \geq (1 - \epsilon_n)\, \Lambda(\eta) + (1 - \epsilon_n) \ < \theta - \eta, \ x > \\ + \epsilon_n \, \Lambda(\lambda_n) + < \theta - \lambda_n, \ \epsilon_n \nabla\Lambda(\lambda_n) > \end{aligned} \tag{2.3.105}$$

where $x \in \partial\Lambda(\eta)$ is as specified in (2.3.100). In the limit $n \to \infty$ by the upper semicontinuity of the convex function $\Lambda(\cdot)$

$$\Lambda(\theta) \geq \Lambda(\eta) + < \theta - \eta, \quad x + y > \tag{2.3.106}$$

(as $\limsup_{n\to\infty} \Lambda(\lambda_n) \leq \Lambda(\eta) < \infty$). Therefore, in particular

$$< \mathbf{v},\ y > \quad < 0 \tag{2.3.107}$$

By comparing (2.3.104) for $\theta = \lambda_n = \eta + \delta_n \mathbf{v}$ with

$$\Lambda(\lambda_n) \geq \Lambda(\eta) + < \lambda_n - \eta, x >= \Lambda(\eta) + \delta_n < \mathbf{v}, x >$$

one obtains

$$\delta_n < \mathbf{v},\ x > \leq\ \delta_n < \mathbf{v},\ \nabla\Lambda(\lambda_n) > \tag{2.3.108}$$

for the same subsequence $\{\lambda_n\}$ as used above. Multiplying by $\frac{\epsilon_n}{\delta_n} > 0$ and taking the limit $n \to \infty$ yields $< \mathbf{v},\ y > \geq\ 0$ in contradiction with (2.3.107). In conclusion, $\partial\Lambda(\eta) = \emptyset$ for any $\eta \in \mathcal{D}_\Lambda \setminus \mathcal{D}_\Lambda^o$. □

Exercises:

2.3.1 Prove that for any $\mu \in M_1(\Sigma)$, the relative entropy $H(\cdot|\mu)$ is the Fenchel-Legendre transform of the function $\Lambda(\cdot)$ defined in (2.3.82).

Hint: Prove first that $\mathcal{D}_{\Lambda^*} = M_1(\Sigma_\mu)$. Then, show that $\nu(a_i) = 0$ and $\nu \in M_1(\Sigma_\mu)$ imply that the value of $\Lambda^*(\nu)$ is obtained by taking $\lambda_i \to -\infty$. Finally, show that $\nu = \nabla\Lambda(\eta)$ when ν is a probability vector with $\Sigma_\nu = \Sigma_\mu = \Sigma$ and

$$\eta_i \stackrel{\Delta}{=} \log\left[\frac{\nu(a_i)}{\mu(a_i)}\right] \quad i = 1, \ldots, |\Sigma| \ . \ . \tag{2.3.109}$$

Conclude that then $\Lambda^*(\nu) =< \eta, \nu > -\Lambda(\eta) = H(\nu|\mu)$.

2.3.2 (a). Use the exponential form of Markov's inequality to prove that for any $C \subset \mathbb{R}^d$ any n and any $\lambda \in \mathbb{R}^d$

$$a_n \log \mu_n(C) \leq -\inf_{y\in C} < \lambda, y > + a_n \Lambda_n(a_n^{-1}\lambda) \ .$$

(b). Assume that $\Lambda(\lambda) = a_n \Lambda_n(a_n^{-1}\lambda)$ for any n (examples where this is true are given in Theorem 2.3.2 and in exercise 2.3.5). Recall the following version of the min-max theorem: let $g(\theta, y)$ be convex and lower semicontinuous in y, concave and upper semicontinuous in θ. Let $C \subset \mathbb{R}^d$ be convex and compact. Then

$$\inf_{y\in C} \sup_\theta g(\theta, y) = \sup_\theta \inf_{y\in C} g(\theta, y)$$

(c.f. [10], pg. 174). Apply this theorem to justify the upper bound

$$a_n \log \mu_n(C) \leq - \sup_{\lambda \in \mathbb{R}^d} \inf_{y \in C} [< \lambda, y > - \Lambda(\lambda)] = - \inf_{y \in C} \Lambda^*(y)$$

for any n and any **convex, compact** set C.

(c). Establish (2.3.64) for all compact sets by applying the above bound. Note that this approach yields a concrete upper bound for any finite n.

(d). Find a compact set B for which

$$\sup_{\lambda \in \mathbb{R}^d} \inf_{y \in B} [< \lambda, y > - \Lambda(\lambda)] < \inf_{y \in B} \Lambda^*(y) \, .$$

2.3.3 Prove that in the assumptions of Theorem 2.3.2, the assumption of lower semicontinuity of Λ may be dropped as it follows from the steepness and i.i.d. structure.

Hint you need to extend exercise 2.2.4 to $\mathbb{R}^d$. The only difficulty is with sequences $\lambda_n \to \lambda$ not along a line, but such that their angle from a line converges to zero. Use the last observation to bound the difference between these two situations in terms of the function G introduced in exercise 2.2.4.

2.3.4 Let $(w_{t_1}, \cdots, w_{t_d})$ be samples of a Brownian motion at $t_1, \ldots, t_d$, i.e. $(w_{t_{j+1}} - w_{t_j})$ is a Normal random variable independent of w_{t_ℓ}, $\ell \leq j$, of variance $(t_{j+1} - t_j)$ and zero mean. Find the rate function for the empirical mean $\hat{S}_n$ of $X_i \stackrel{\Delta}{=} (w^i_{t_1}, \ldots, w^i_{t_d})$ where $w^i_{t_j}$, $i = 1, \ldots, n$ are samples of independent Brownian motions at instances t_j.

Remark: Note that the law of $\hat{S}_n$ is the same as that of $\frac{1}{\sqrt{n}}(w_{t_1}, \ldots, w_{t_d})$, and compare to Schilder's Theorem which is presented in Section ??.

2.3.5 Let X_j be i.i.d. random variables over $\mathbb{R}^d$ with a steep logarithmic moment generating function Λ such that $0 \in \mathcal{D}^o_\Lambda$. Let $N(t)$ be a Poisson process of unit rate which is independent of the X_j variables and consider the random variables

$$\hat{S}_n \stackrel{\Delta}{=} \frac{1}{n} \sum_{j=1}^{N(n)} X_j \, .$$

Prove that the family of laws μ_n corresponding to $\hat{S}_n$ satisfies a large deviations principle with the rate function being the Fenchel-Legendre transform of $e^{\Lambda(\lambda)} - 1$.

Hint: You can apply Theorem 2.3.2 as $N(n) = \sum_{j=1}^{n} N_j$ where N_j are i.i.d. Poisson(1) random variables.

2.3.6 Let $N(n)$ be a sequence of integer valued random variables whose logarithmic moment generating functions Λ_n satisfy the Assumption 2.3.1. Let X_j be i.i.d. random variables over $\mathbb{R}^d$ with *finite everywhere* logarithmic moment generating function Λ_X and let μ_n denotes the law of

$$Z_n \stackrel{\Delta}{=} a_n \sum_{j=1}^{N(n)} X_j \ .$$

Prove that if the conditions of Lemma 2.3.2 hold for the convex function $\Lambda(\Lambda_X(\lambda))$ then μ_n satisfies a large deviations principle governed by the Fenchel-Legendre transform of this function.

2.3.7 For any $\delta > 0$ let $Z_{n,\delta} = Z_n + \sqrt{\delta a_n} V$ where V is a standard multivariate Normal random variable.

(a). Prove that when Assumption 2.3.1 holds for Z_n it also holds for $Z_{n,\delta}$ with the limiting logarithmic moment generating function $\Lambda_\delta(\lambda) = \Lambda(\lambda) + \frac{\delta}{2}|\lambda|^2$.

(b). Show that for any $x \in \mathbb{R}^d$ the value of the Fenchel-Legendre transform of Λ_δ does not exceed $\Lambda^*(x)$.

(c). Prove that if $\lim_{n\to\infty} E(Z_n) = \overline{x} \in \mathbb{R}^d$ exists then $\Lambda(\lambda) \geq < \lambda, \overline{x} >$ for any $\lambda \in \mathbb{R}^d$. Conclude that if in addition Λ is finite and differentiable everywhere then $\mathcal{F}_\delta = \mathbb{R}^d$ (where $\mathcal{F}_\delta \stackrel{\Delta}{=} \{x : x = \nabla\Lambda_\delta(\lambda) \quad \text{for some} \quad \lambda \in \mathbb{R}^d\}$).

(d). By applying Theorem 2.3.1 for $Z_{n,\delta}$ and (a)-(c) above deduce that for any $x \in \mathbb{R}^d$ and any $\epsilon > 0$

$$\liminf_{n\to\infty} a_n \log \text{Prob}(Z_{n,\delta} \in B_{x,\epsilon/2}) \geq -\Lambda^*(x) \tag{2.3.110}$$

(e). Prove that

$$\limsup_{n\to\infty} a_n \log \text{Prob}(\sqrt{\delta a_n}|V| \geq \epsilon/2) \leq -\frac{\epsilon^2}{8\delta} \tag{2.3.111}$$

(f). Prove that

$$\text{Prob}(Z_n \in B_{x,\epsilon}) \geq \text{Prob}(Z_{n,\delta} \in B_{x,\epsilon/2}) - \text{Prob}(\sqrt{\delta a_n}|V| \geq \epsilon/2) \tag{2.3.112}$$

and by combining (2.3.110), (2.3.111) and (2.3.112) (for $n \to \infty$ and then $\delta \to 0$) conclude that the large deviations lower bound holds for the laws μ_n corresponding to Z_n.

(g). Deduce now by part (a) of Theorem 2.3.1 that when Assumption 2.3.1 holds with Λ which is finite and differentiable everywhere and when moreover $\lim_{n\to\infty} E(Z_n)$ exists then μ_n satisfy a large deviations principle with rate function Λ^*.

Remark: This may serve for example as an alternative derivation of Cramer's theorem which avoids the convex analysis Lemma 2.3.2.

2.3.8 Let $X_1, \ldots, X_n, \ldots$ be a real-valued, zero mean, stationary Gaussian process with covariance sequence $R_i \stackrel{\Delta}{=} E(X_n X_{n+i})$. Suppose the process has a finite power P defined via $P \stackrel{\Delta}{=} \lim_{n\to\infty} \sum_{i=-n}^{n} R_i(1 - \frac{|i|}{n})$. Let μ_n be the law of the empirical mean $\hat{S}_n$ of the first n samples of this process. Prove that μ_n satisfy a large deviations principle controlled by the good rate function $\Lambda^*(x) = \frac{x^2}{2P}$.

2.3.9 Again, let $X_1, \ldots, X_n, \ldots$ be a real-valued, zero mean, stationary Gaussian process with covariance sequence $R_i \stackrel{\Delta}{=} E(X_n X_{n+i})$. Assume that this covariance sequence is absolutely summable and let $S(\omega) > 0$ denote its Fourier transform. Consider the empirical covariances

$$Z_n^j = \frac{1}{n} \sum_{i=1}^{n-j} X_i X_{i+j}$$

for $j = 0, \ldots, d-1$. Let $Z_n \in \mathbb{R}^d$ be the empirical covariance vector composed of $\{Z_n^j\}$ and Λ_n the corresponding logarithmic moment generating functions.

(a). Verify that

$$\Lambda_n(n\theta) = -\frac{1}{2} \sum_{i=1}^{n} \log[1 - \lambda_i(\Theta \mathbf{R})] \, .$$

Here $\lambda_i(\Theta\mathbf{R})$ is the i-th eigenvalue of the product of the covariance matrix $\mathbf{R}$ and the matrix Θ where $\Theta(j,k) = \theta_{j-k}$ for all $j \in \{k, \ldots, k+d-1\}$ and is zero otherwise.

(b). Assume that

$$\lim_{n\to\infty} -\frac{1}{2n} \sum_{i=1}^{n} \log[1 - \lambda_i(\Theta\mathbf{R})] = -\frac{1}{4\pi} \int_0^{2\pi} \log[1 - S(\omega) \sum_{k=0}^{d-1} \theta_k e^{-i\omega k}] d\omega \stackrel{\Delta}{=} \Lambda(\theta)$$

(this identity indeed holds by the limiting distribution of near Toeplitz matrices, see [16]). Prove that the empirical covariance vectors Z_n satisfy a large deviations principle controlled by the Fenchel-Legendre transform of the function Λ defined above.

2.4 Large deviations of Markov chains over finite alphabets

The results of Section 2.1 are extended in this section to random variables $X_1, \ldots, X_n$ which take values in the finite alphabet $\Sigma = \{a_1, \ldots a_{|\Sigma|}\}$, with a Markov structure instead of an i.i.d. structure. Although most of the results may be derived by the method of types presented in Section 2.1, the combinatorics involved are quite cumbersome (for more details about this approach consider exercise 2.4.7). Thus, an alternative derivation of these results via an application of Theorem 2.3.1 is adopted here. Without loss of generality, identify Σ with the set $\{1, \ldots |\Sigma|\}$ so that $a_i = i$.

Let $\Pi = \{\pi(i,j)\}_{i,j=1\ldots|\Sigma|}$ be a stochastic matrix, i.e. a matrix whose elements are non-negative and such that each row-sum is one. P_x^π denotes the Markov probability measure associated with the transition probability Π and initial state $x \in \Sigma$. Specifically,

$$P_x^\pi(X_1 = x_1, \ldots, X_n = x_n) = \pi(x, x_1) \prod_{i=1}^{n-1} \pi(x_i, x_{i+1}) . \tag{2.4.113}$$

A matrix $\mathbf{B}$ with nonnegative entries is called *irreducible*, if for any pair of indices i, j there exists an $m(i,j)$ such that $B^{m(i,j)}(i,j) > 0$. where $\mathbf{B}^m$ denotes the usual product of matrices. This property is equivalent to the condition that one may find for each i, j a sequence of indices $i_1, \ldots, i_m$ such that $i_1 = i$, $i_m = j$ and $B(i_k, i_{k+1}) > 0$ for all $k = 1, \ldots, m-1$. The following theorem describes basic properties of irreducible matrices.

Theorem 2.4.1 (Perron-Frobenius) *Let* $\mathbf{B} = \{B(i,j)\}_{i,j=1}^{|\Sigma|}$ *be an irreducible matrix. Then there exists an eigenvalue* ρ *(called the Perron-Frobenius eigenvalue) such that*

(a). $\rho > 0$ *is real.*

(b). There exist right and left eigenvectors corresponding to the eigenvalue ρ *which are strictly positive, i.e. there exist vectors* μ, ϑ *with* $\mu_i, \vartheta_i > 0$ *for all* i *(this is denoted in the sequel by* $\mu >> 0, \vartheta >> 0$*) such that*

$$\sum_{j=1}^{|\Sigma|} B(i,j)\vartheta_j = \rho\vartheta_i \tag{2.4.114}$$

$$\sum_{i=1}^{|\Sigma|} \mu_i B(i,j) = \rho\mu_j \tag{2.4.115}$$

(c). For any eigenvalue λ of $\mathbf{B}$, $|\lambda| \leq \rho$.

(d). The right and left eigenvectors μ, ϑ corresponding to the eigenvalue ρ are unique up to a constant multiple.

(e). Let ϕ be any strictly positive vector, then for any $i \in \Sigma$

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(i,j)\, \phi_j \right] = \lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(j,i)\, \phi_j \right] = \log \rho \tag{2.4.116}$$

Proof: Parts *(a)÷(d)* are stated for example in [27], Theorem 1.5. To prove part *(e)*, let $\alpha \triangleq \sup_i \vartheta_i$, $\beta \triangleq \inf_i \vartheta_i > 0$ and $\gamma \triangleq \inf_j \phi_j > 0$, $\delta = \sup_j \phi_j$ (where ϑ is the right eigenvector corresponding to ρ above). Then, for all $i, j \in \Sigma$,

$$\frac{\delta}{\beta} B^n(i,j)\, \vartheta_j \geq B^n(i,j)\, \phi_j \geq \frac{\gamma}{\alpha} B^n(i,j)\, \vartheta_j \tag{2.4.117}$$

Therefore,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(i,j)\, \phi_j \right] &= \lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(i,j)\, \vartheta_j \right] \\ = \lim_{n\to\infty} \frac{1}{n} \log (\rho^n\, \vartheta_i) &= \log \rho \ . \end{aligned} \tag{2.4.118}$$

A similar argument leads to

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{j=1}^{|\Sigma|} B^n(j,i) \phi_j \right] = \log \rho \, . \tag{2.4.119}$$

□

2.4.1 Cramer's theorem for Markov additive processes

Let $g : \Sigma \to \mathbb{R}^d$ be a given deterministic function. The large deviations of the empirical g-mean

$$Z_n = \frac{1}{n} \sum_{k=1}^{n} g(X_k) \tag{2.4.120}$$

are the subject of this section (for the extension to random functions see exercise 2.4.1). If X_k were independent, Cramer's Theorem 2.1.1 applies and then the large deviations principle is governed

by the Legendre transform of the logarithmic moment generating function. Theorem 2.3.1 hints that the rate function may still be expressed in terms of a Legendre transform even in the current dependent case where the X_k possess a Markov structure.

In order to find the proper function which replaces the logarithmic moment generating function $\Lambda(\lambda)$ associate with any $\lambda \in \mathbb{R}^d$ a non-negative matrix Π_λ via

$$\pi_\lambda(i,j) = \pi(i,j)\, e^{<\lambda, g(j)>} \quad i,j \in \Sigma \ . \tag{2.4.121}$$

When Π is irreducible it follows that Π_λ are also irreducible matrices since $e^{<\lambda,g(j)>}$ is positive. Let $\rho(\Pi_\lambda)$ denotes the Perron-Frobenius eigenvalue of Π_λ then $\log\, \rho(\Pi_\lambda)$ plays the role of the logarithmic moment generating function $\Lambda(\lambda)$. Specifically, the following analog of Theorem 2.3.2 holds.

Theorem 2.4.2 *Assume Π is irreducible and define*

$$I(z) \stackrel{\Delta}{=} \sup_{\lambda \in \mathbb{R}^d} \{< \lambda, z > - \log\, \rho(\Pi_\lambda)\} \tag{2.4.122}$$

Then, $I(\cdot)$ is a good, convex, rate function which controls the large deviations of the empirical g-means $\{Z_n\}$, i.e. for any measurable set $\Gamma \subseteq \mathbb{R}^d$, and any initial state $x \in \Sigma$,

$$-\inf_{z \in \Gamma^o} I(z) \leq \liminf_{n\to\infty} \frac{1}{n} \log P_x^\pi\left(\{Z_n \in \Gamma\}\right) \leq \limsup_{n\to\infty} \frac{1}{n} \log P_x^\pi(\{Z_n \in \Gamma\}) \leq -\inf_{z \in \overline{\Gamma}} I(z) \tag{2.4.123}$$

Proof: Define

$$\Lambda_n(\lambda) \stackrel{\Delta}{=} \log\, E_x^\pi \left[e^{<\lambda, Z_n>}\right] \tag{2.4.124}$$

In view of Theorem 2.3.1, it is enough to check that the limit

$$\Lambda(\lambda) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} \Lambda_n(n\lambda) = \lim_{n\to\infty} \frac{1}{n} \log E_x^\pi \left[e^{n<\lambda, Z_n>}\right] \tag{2.4.125}$$

exists for all $\lambda \in \mathbb{R}^d$, is differentiable, and that $\Lambda(\lambda) = \log \rho(\Pi_\lambda)$. Note that

$$\begin{aligned}
\Lambda_n(n\lambda) &= \log E_x^\pi \left[e^{<\lambda, \sum_{k=1}^n g(X_k)>}\right] \\
&= \log \sum_{x_1,\ldots,x_n} P_x^\pi(X_1 = x_1, \ldots, X_n = x_n) \prod_{k=1}^n e^{<\lambda, g(x_k)>} \\
&= \log \sum_{x_1,\ldots,x_n} \pi(x,x_1) e^{<\lambda,g(x_1)>} \cdots \pi(x_{n-1},x_n) e^{<\lambda,g(x_n)>} \\
&= \log \sum_{j=1}^{|\Sigma|} (\Pi_\lambda)^n(x,j)
\end{aligned} \tag{2.4.126}$$

Since Π_λ is an irreducible matrix, part *(e)* of the Perron-Frobenius theorem yields (for $\phi_j = 1$)

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \Lambda_n(n\lambda) = \log \rho(\Pi_\lambda) \tag{2.4.127}$$

Moreover, since $|\Sigma|$ is finite it is clear that $\rho(\Pi_\lambda)$, being an isolated root of the characteristic equation for the matrix Π_λ is differentiable with respect to λ (see [14] for details). Therefore, Theorem 2.3.1 may be applied to complete the proof. □

Remark: The above proof relies on two properties of the Markov chain – namely, part *(e)* of the Perron-Frobenius theorem and the differentiability of $\rho(\Pi_\lambda)$ with respect to λ. Thus, Theorem 2.4.2 holds as long as the Markov chain has these two properties. In particular, the finiteness of Σ is not crucial and indeed a large deviations principle for a general Markov chain set-up is presented in Section ??.

Exercises:

2.4.1 Assume that $X_1, \ldots, X_n$ have the joint law P_x^π where Π is an irreducible stochastic matrix. Consider the empirical means

$$Z_n = \frac{1}{n} \sum_{k=1}^{n} Y_k$$

where the conditional law of Y_k when $X_k = j$ is $\mu_j \in M_1(\mathbb{R}^d)$ and for any given realization $\{X_k\}_{k=1}^n$ of the Markov chain states the variables Y_k are conditionally independent. Suppose that the logarithmic moment generating functions Λ_j associated with μ_j are finite everywhere (for all $j \in \Sigma$). Prove that Theorem 2.4.2 holds in this case where now

$$\pi_\lambda(i,j) \stackrel{\Delta}{=} \pi(i,j)\, e^{\Lambda_j(\lambda)} \quad i,j \in \Sigma \ .$$

2.4.2 Sanov's theorem for the empirical measure of Markov chains

A particularly important application of Theorem 2.4.2 above yields the large deviations principle satisfied by the empirical measure of Markov chains. Namely, define $L_n^{\mathbf{X}}(i) = \frac{1}{n} \sum_{j=1}^n \delta_i(X_j)$, where

$$\delta_i(x) \stackrel{\Delta}{=} \begin{cases} 1 & i = x \\ 0 & \text{otherwise} \end{cases} .$$

For $i = 1, \ldots, |\Sigma|$, let $a_n(i) = E_x^\pi[L_n^{\mathbf{X}}(i)] = \frac{1}{n} \sum_{j=1}^n \pi^j(x,i)$. Then, $a_n \to \mu$ as $n \to \infty$ where μ is a unique properly normalized left eigenvector of Π (since $|a_n(\Pi - \mathbf{I})(j)| = \frac{1}{n}|\pi^{n+1}(x,j) - \pi(x,j)| \le \frac{2}{n}$).

Further, by Chebychev's inequality, $L_n^{\mathbf{X}} \to \mu$ as $n \to \infty$ and therefore, $L_n^{\mathbf{X}}$ is a good candidate for a large deviations statement on $M_1(\Sigma)$.

It is clear that $L_n^{\mathbf{X}}$ fits into the framework of Section 2.4.1 with $g(x) = (\delta_1(x), \ldots, \delta_{|\Sigma|}(x))$. Therefore, by Theorem 2.4.2, a large deviations principle follows with the rate function

$$I(q) = \sup_{\lambda \in \mathbb{R}^d} (< \lambda, q > - \log \rho(\Pi_\lambda)) , \tag{2.4.128}$$

where here $\pi_\lambda(i,j) \stackrel{\Delta}{=} \pi(i,j)\, e^{\lambda_j}$ and $q \in M_1(\Sigma)$. The following alternative characterization of $I(q)$ is sometimes more useful.

Theorem 2.4.3

$$I(q) = J(q) \stackrel{\Delta}{=} \begin{cases} \infty & q \notin M_1(\Sigma) \\ \sup\limits_{\mathbf{u}>>0} \sum\limits_{j=1}^{|\Sigma|} q_j \log \left[\frac{u_i}{(\mathbf{u}\Pi)_j} \right] , & q \in M_1(\Sigma) \end{cases} \tag{2.4.129}$$

Remarks: This identity actually holds also for non-stochastic matrices (see exercise 2.4.3). In the i.i.d. set up the rows of Π are identical and then $J(q)$ is merely the relative entropy $H(q|\pi(1,\cdot))$ (see exercise 2.4.2).

Proof: Since $M_1(\Sigma)$ is a closed subset of $\mathbb{R}^{|\Sigma|}$ its complement $M_1(\Sigma)^c$ is an open set. Further, $L_n^{\mathbf{X}} \in M_1(\Sigma)$, for any n and any realization $\mathbf{X}$. Therefore, by the large deviations lower bound (2.4.123)

$$-\infty = \frac{1}{n} \log P_x^\pi(\{L_n^{\mathbf{X}} \in M_1(\Sigma)^c\}) \geq - \inf_{q \notin M_1(\Sigma)} I(q) , \tag{2.4.130}$$

i.e., $I(q) = \infty$ for any $q \notin M_1(\Sigma)$.

Fix $q \in M_1(\Sigma)$, $\mathbf{u} >> 0$ and set $\lambda_j = \log \left[\frac{u_i}{(\mathbf{u}\Pi)_j} \right]$ (since $\mathbf{u} >> 0$ and Π is irreducible, it follows that $\mathbf{u}\Pi >> 0$). Observe that $\mathbf{u}\Pi_\lambda = \mathbf{u}$ and thus $\rho(\Pi_\lambda) = 1$ by part *(e)* of the Perron-Frobenius theorem (with $\phi_i = u_i > 0$). Therefore,

$$I(q) \geq \sum_{j=1}^{|\Sigma|} q_j \log \left[\frac{u_j}{(\mathbf{u}\Pi)_j} \right] - \log 1$$

Since $\mathbf{u} >> 0$ is arbitrary this inequality implies that $I(q) \geq J(q)$.

To establish the reverse inequality, fix an arbitrary vector $\lambda \in \mathbb{R}^{|\Sigma|}$ and let $\alpha > \rho(\Pi_\lambda)$ be any upper bound on $\rho(\Pi_\lambda)$. Define

$$u_j^* \triangleq \sum_{n=0}^{\infty} \alpha^{-n} \sum_{i=1}^{|\Sigma|} (\Pi_\lambda)^n(i,j) < \infty, \quad j = 1, \ldots, |\Sigma| \ , \tag{2.4.131}$$

where the finiteness of u_j^* is a direct consequence of part *(e)* of the Perron-Frobenius theorem. Moreover, $\mathbf{u}^*\Pi_\lambda = \alpha(\mathbf{u}^* - 1)$ (where 1 denotes the all ones vector) and $\mathbf{u}^* >> 0$. Thus, by the definition of Π_λ,

$$\begin{aligned} < \lambda, q > \ + \ \sum_{j=1}^{|\Sigma|} q_j \log \frac{(\mathbf{u}^*\Pi)_j}{u_j^*} &= \sum_{j=1}^{|\Sigma|} q_j \log \frac{(\mathbf{u}^*\Pi_\lambda)_j}{u_j^*} \\ = \ \sum_{j=1}^{|\Sigma|} q_j \log \frac{\alpha(\mathbf{u}^* - 1)_j}{u_j^*} &= \log \alpha + \sum_{j=1}^{|\Sigma|} q_j \log \left(1 - \frac{1}{u_j^*}\right) < \log \alpha \end{aligned} \tag{2.4.132}$$

Thus, $(< \lambda, q > - \log \alpha) < \sum_{j=1}^{|\Sigma|} q_j \log \frac{u_j^*}{(\mathbf{u}^*\Pi)_j}$. Take $\alpha \downarrow \rho(\Pi_\lambda)$ to deduce that

$$< \lambda, q > - \log \rho(\Pi_\lambda) \leq \sup_{\mathbf{u} >> 0} \sum_{j=1}^{|\Sigma|} q_j \log \frac{u_j}{(\mathbf{u}\Pi)_j} \ = \ J(q) \tag{2.4.133}$$

and since λ is arbitrary, also $I(q) \leq J(q)$ and the proof is complete. □

Exercises:

2.4.2 Suppose $\pi(i,j) = \mu(j)$, $i,j \in \Sigma$ where $\mu \in M_1(\Sigma)$. Prove that then $J(\cdot) = H(\cdot|\mu)$ (the relative entropy with respect to μ) while $I(\cdot)$ is the Fenchel-Legendre transform of the moment generating function Λ of (2.3.82). Thus, Theorem 2.4.3 is the natural extension of exercise 2.3.1 to the Markov set-up.

2.4.3 (a). Show that (2.4.129) holds for any non-negative irreducible matrix Π (not necessarily stochastic).

Hint: Let $\phi(i) = \sum_j \pi(i,j)$. Clearly, $\phi >> 0$, and thus the matrix Π^* given by $\pi^*(i,j) = \pi(i,j)/\phi(i)$ is stochastic. Prove now that $J_{\Pi^*}(q) = J_\Pi(q) + \sum_j q_j \log \phi(j)$ for any $q \in \mathbb{R}^{|\Sigma|}$, and likewise $I_{\Pi^*}(q) = I_\Pi(q) + \sum_j q_j \log \phi(j)$ (where J_Π and I_Π denote the rate functions J and I associated with the matrix Π via (2.4.129) and (2.4.128) respectively).

(b). Show that for any irreducible, non-negative matrix Π

$$\log \rho(\Pi) = \sup_{\nu \in M_1(\Sigma)} \{-J(\nu)\} \tag{2.4.134}$$

This characterization of the spectral radius of non-negative matrices is useful when looking for tight bounds (for an alternative characterization see exercise 2.4.5).

2.4.3 Sanov's theorem for the pair empirical measure of Markov chains

The large deviations principle for the empirical measure of a Markov chain is still in the form of an optimization problem. Moreover the nice interpretation in terms of entropy (recall Section 2.1.1 where the i.i.d. case is presented) has disappeared. It is interesting to note that by considering a somewhat different random variable, from which the large deviations for $L_n^{\mathbf{X}}$ may be recovered (see exercise 2.4.4), one is also able to get a large deviations with a rate function which is an appropriate relative entropy.

Consider the space $\Sigma^{(2)} \triangleq \Sigma \times \Sigma$, which corresponds to consecutive pairs of elements from the sequence $\mathbf{X}$. Note that by considering the pairs formed by $X_1, \ldots, X_n$, i.e. the sequence $X_1X_2, X_2X_3, \ldots, X_iX_{i+1}, \ldots, X_{n-1}X_n$, one recovers a Markov chain with state space $\Sigma^{(2)}$ and transition matrix $\Pi^{(2)}$ specified via

$$\pi^{(2)}(k \times \ell,\ i \times j) = \delta_\ell(i)\, \pi(i,j) \tag{2.4.135}$$

For simplicity assume throughout this section that Π is strictly positive (i.e. $\pi(i,j) > 0$ for all i,j). Then, $\Pi^{(2)}$ is an irreducible transition matrix, and therefore the results of Section 2.4.2 may be applied to find the large deviations rate function $I^{(2)}(q)$ associated with the pair empirical measures

$$L_n^{\mathbf{X},(2)}(y) \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_y(X_{i-1}, X_i), \quad y \in \Sigma^{(2)} \ . \tag{2.4.136}$$

Note that $L_n^{\mathbf{X},(2)} \in M_1(\Sigma^{(2)})$ and therefore $I^{(2)}(\cdot)$ is a good, convex, rate function over this space. The next theorem characterizes $I^{(2)}(\cdot)$ as an appropriate relative entropy. The following definitions are needed for that purpose. For any $q \in M_1(\Sigma^{(2)})$, let $q(i) \triangleq \sum_{j=1}^{|\Sigma|} q(i,j)$ be its marginal and when $q(i) > 0$ let $q(j|i) \triangleq \frac{q(i,j)}{q(i)}$. A measure $q \in M_1(\Sigma^{(2)})$ is called **shift invariant** if $q(i) = \sum_{k=1}^{|\Sigma|} q(k,i)$ for all i (i.e., both marginals of q are identical).

Theorem 2.4.4 *Assume Π is strictly positive. Then for any $q \in M_1(\Sigma^{(2)})$,*

$$I^{(2)}(q) = \begin{cases} \sum_i q(i) H(q(\cdot|i) \,|\, \pi(i,\cdot)), & q \ \textit{shift invariant} \\ \infty & \textit{otherwise} \end{cases} \tag{2.4.137}$$

where $H(\cdot|\cdot)$ is the relative entropy function defined in Section 2.1.1. Specifically,

$$H(q(\cdot|i)|\pi(i,\cdot)) = \sum_{j=1}^{|\Sigma|} q(j|i) \log \frac{q(j|i)}{\pi(i,j)}. \tag{2.4.138}$$

Remarks: When Π is not strictly positive (but is irreducible) the theorem still applies with $\Sigma^{(2)}$ replaced by $\Sigma_\Pi \triangleq \{(i,j) : \pi(i,j) > 0\}$, and an almost identical proof. The above representation of $I^{(2)}(q)$ is useful for example in characterizing the spectral radius of non-negative matrices (see exercise 2.4.5) and in establishing the analog of Sanov's theorem for *time weighted* empirical measures (see exercise 2.4.6). It is also useful because bounds on the relative entropy are readily available and may be used to obtain bounds on the rate function.

Proof: By Theorem 2.4.3

$$\begin{aligned} I^{(2)}(q) &= \sup_{\mathbf{u}>>0} \sum_{j=1}^{|\Sigma|}\sum_{i=1}^{|\Sigma|} q(i,j) \log \frac{u(i,j)}{(\mathbf{u}\Pi^{(2)})(i,j)} \\ &= \sup_{\mathbf{u}>>0} \sum_{j=1}^{|\Sigma|}\sum_{i=1}^{|\Sigma|} q(i,j) \log \frac{u(i,j)}{\left[\sum_k u(k,i)\right]\pi(i,j)} \end{aligned} \tag{2.4.139}$$

where the last equality follows by (2.4.135).

Assume first that q is not shift invariant. Then, $q(j_0) < \sum_k q(k, j_0)$ for some j_0. For $\mathbf{u}$ such that $u(\cdot, j) = 1$ when $j \neq j_0$ and $u(\cdot, j_0) = e^\alpha$,

$$\sum_{j=1}^{|\Sigma|}\sum_{i=1}^{|\Sigma|} q(i,j) \log \left[\frac{u(i,j)}{[\sum_k u(k,i)]\,\pi(i,j)}\right] = \sum_{j=1}^{|\Sigma|}\sum_{i=1}^{|\Sigma|} q(i,j) \log \left[\frac{u(1,j)}{|\Sigma| u(1,i)\pi(i,j)}\right]$$

$$= -\sum_{j=1}^{|\Sigma|}\sum_{i=1}^{|\Sigma|} q(i,j) \log\{|\Sigma|\pi(i,j)\} + \alpha\left[\sum_{i=1}^{|\Sigma|} q(i,j_0) - q(j_0)\right] \tag{2.4.140}$$

which implies, by considering $\alpha \to \infty$, that $I^{(2)}(q) = \infty$.

Finally, when q is shift invariant then for any $\mathbf{u} >> 0$

$$\sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|} q(i,j) \log \left[\frac{\sum_k u(k,i)q(j)}{\sum_k u(k,j)\, q(i)}\right] = 0. \tag{2.4.141}$$

Let $u(i|j) = u(i,j)/\sum_k u(k,j)$ and $q(i|j) = q(i,j)/q(j)$, i.e., $q(i|j) = q(i,j)/\sum_k q(k,j)$ (since q is

shift invariant). Now, by (2.4.139) and (2.4.141)

$$I^{(2)}(q) - \sum_{i=1}^{|\Sigma|} q(i)\, H(q(\cdot|i)|\pi(i,\cdot)) = \sup_{u>>0} \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|} q(i,j)\, \log \frac{u(i,j)q(i)}{[\sum_k u(k,i)]q(i,j)}$$

$$= \sup_{u>>0} \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|} q(i,j) \log \frac{u(i|j)}{q(i|j)} = \sup_{u>>0} \left\{ -\sum_{j=1}^{|\Sigma|} q(j) H(q(\cdot|j)|u(\cdot|j)) \right\} \qquad (2.4.142)$$

Since $H(q(\cdot|j)|u(\cdot|j)) \geq 0$ it follows that $I^{(2)}(q) \leq \sum_i q(i)\, H(q(\cdot|i)|\pi(i,\cdot))$ with equality whenever $q >> 0$ (by the choice $u = q$). The proof is complete for q which is not strictly positive by considering a sequence $u_n >> 0$ such that $u_n \to q$ (so $q(j)H(q(\cdot|j)|u_n(\cdot|j)) \to 0$ for each j). □

Exercises:

2.4.4 Prove that for any strictly positive stochastic matrix Π

$$J(\nu) = \inf_{\{q:\sum_i q(i,\cdot)=\nu\}} I^{(2)}(q) \qquad (2.4.143)$$

where $J(\cdot)$ is the rate function defined in (2.4.129) while $I^{(2)}(\cdot)$ is as specified in (2.4.137).

Hint: There is no need to prove the above identity directly. Instead observe that the empirical measure $L_n^{\mathbf{X}}$ belongs to a set A iff $L_n^{\mathbf{X},(2)} \in \{q : \sum_i q(i,\cdot) \in A\}$ (where the initial condition $X_0 = x$ is equivalent to any initial condition $X_{-1}X_0 = (i,x)$ for the $\Pi^{(2)}$ chain). As the projection of any measure $q \in M_1(\Sigma^{(2)})$ onto its marginal $\nu \in M_1(\Sigma)$ is continuous and $I^{(2)}(\cdot)$ controls the large deviations of $L_n^{\mathbf{X},(2)}$ deduce that the right side of (2.4.143) is a rate function governing the large deviations of $L_n^{\mathbf{X}}$. Conclude by proving the uniqueness of such a function and applying Theorem 2.4.3.

2.4.5 (a). Extend the validity of the identity (2.4.143) to any irreducible non-negative matrix Π.

Hint: First extend Theorem 2.4.4 to any irreducible stochastic matrix Π by replacing $\Sigma^{(2)}$ with Σ_Π (see the remark following the statement of this theorem). Then, for any irreducible, non-negative matrix Π consider Π^* defined in exercise 2.4.3 and verify that $I^{(2)}_{\Pi^*}(q) = I^{(2)}_{\Pi}(q) + \sum_i q(i) \log \phi(i)$.
(b). Deduce by applying the identities (2.4.134) and (2.4.143) that for any non-negative irreducible matrix Π

$$-\log \rho(\Pi) = \inf_{q \in M_1(\Sigma_\Pi)} I^{(2)}(q) = \inf_{q \in M_1(\Sigma_\Pi) \text{ shift invariant}} \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|} q(i,j)\, \log \frac{q(j|i)}{\pi(i,j)}\,. \qquad (2.4.144)$$

This is Varadhan's characterization of the spectral radius of non-negative irreducible matrices which is extremely useful for many applications.

2.4.6 Assume that $X_1, \ldots, X_n$ have the joint law P_x^π where Π is an irreducible stochastic matrix. Let $T_1, \ldots, T_n$ be a sequence of random variables over $\mathcal{T} \stackrel{\Delta}{=} \{1, 2, \ldots, \ell\}$ which are conditionally independent given any realization $\{X_k\}_{k=1}^n$ while $\text{Prob}(T_k = t | X_{k-1} = i) = p(i,t)$, $i \in \Sigma$, $t \in \mathcal{T}$. Construct the partial sums $S_m = \sum_{k=1}^m T_k$ and let K_n be the stopping time where S_m first hits or exceeds the integer value n. The **time weighted empirical measures** $L_n^{\mathbf{X},\mathbf{T}}$ are defined via

$$L_n^{\mathbf{X},\mathbf{T}}(j) = \frac{1}{n}[\sum_{k=1}^{K_n - 1} T_k \delta_j(X_k) + (n - S_{K_n - 1})\delta_j(X_{K_n})] \, .$$

(a). Suppose that $p(i,\ell) > 0$ for any $i \in \Sigma$. Prove that $L_n^{\mathbf{X},\mathbf{T}}$ satisfies a large deviations principle with rate function

$$J(\nu) \stackrel{\Delta}{=} \inf_{A_\nu} \left\{ \sum_{i=1}^{|\Sigma|} a(i) \sum_{j=1}^{|\Sigma|} \sum_{t=1}^{\ell} q(j,t|i) \log \frac{q(j,t|i)}{\pi(i,j)p(i,t)} \right\}$$

where

$$A_\nu \stackrel{\Delta}{=} \{q(\cdot|i) \in M_1(\Sigma \times \mathcal{T}), a(i) \geq 0 : \sum_{i=1}^{|\Sigma|} a(i) \sum_{t=1}^{\ell} q(j,t|i) = a(j), \sum_{i=1}^{|\Sigma|} a(i) \sum_{t=1}^{\ell} t q(j,t|i) = \nu(j), \forall j \in \Sigma\} \, .$$

Hint: Interpret the event $\{T_k = t\}$ as if the Markov chain freezes in its current state X_k for t time units and observe that $L_n^{\mathbf{X},\mathbf{T}}$ is merely the standard empirical measure in this new time scale. Consider the Markov chain whose state space $\Sigma \times \mathcal{T}$ consists of the original states and the future time spans in which state changes are still forbidden (starting at the initial state $(x,1)$). Show that the transition from (i,s) to (j,t) in this chain has probability $\pi(i,j)p(i,t)$ when $s = 1$ and $\delta_i(j)\delta_{s-1}(t)$ otherwise. Apply Theorem 2.4.4 to the pair empirical measure of this chain $L_n^{\mathbf{X},(2)}$. Finally, observe that $L_n^{\mathbf{X},\mathbf{T}}(j) = \sum_{i,s,t} \; t L_n^{\mathbf{X},(2)}((i,s),(j,t))$ for any $j \in \Sigma$, and apply a "contraction argument" of the type hinted about in exercise 2.4.4.

(b). Suppose that $p(i,t) = p(t)$ and $\pi(i,j) = \mu(j)$. Prove that now the rate function for $L_n^{\mathbf{X},\mathbf{T}}$ is

$$J(\nu) = \inf_{\{q : E_q[T_1 \delta_j(X_1)] = \nu(j) E_q[T_1], \, \forall j\}} \frac{H(q|\mu \times p)}{E_q(T_1)}$$

where $E_q(T_1) = \sum_{j,t} \; t q(j,t)$ and $q \in M_1(\Sigma \times \mathcal{T})$.

2.4.7 (a). Prove that

$$\frac{1}{n}\log P_x^\pi(X_1=x_1,\ldots,X_n=x_n)=\sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|}L_n^{\mathbf{x},(2)}(i,j)\log\pi(i,j)$$

for any sequence $\mathbf{x}=(x_1,\ldots,x_n)\in\Sigma^n$ of non-zero P_x^π probability.

(b). Let

$$\mathcal{L}_n\stackrel{\Delta}{=}\{q: q=L_n^{\mathbf{x},(2)},\ P_x^\pi(X_1=x_1,\ldots,X_n=x_n)>0\ \text{ for some } \mathbf{x}\in\Sigma^n\}$$

be the set of possible types of pairs of states of the Markov chain. Prove that $\mathcal{L}_n\subset M_1(\Sigma_\Pi)$ and $|\mathcal{L}_n|\le(n+1)^{|\Sigma|^2}$.

(c). Let $T(q)$ be the type class of $q\in\mathcal{L}_n$, namely the set of sequences $\mathbf{x}$ of positive P_x^π probability for which $L_n^{\mathbf{x},(2)}=q$ and let $H(q)\stackrel{\Delta}{=}\sum_{i,j}q(i,j)\log q(j|i)$. Suppose that for any $q\in\mathcal{L}_n$

$$(n+1)^{-(|\Sigma|^2+|\Sigma|)}e^{nH(q)}\le|T(q)|\le e^{nH(q)} \tag{2.4.145}$$

and moreover that

$$\lim_{n\to\infty}d_V(q,\mathcal{L}_n)=0\quad\forall q\in M_1(\Sigma_\Pi),\ \ q\ \text{ shift invariant}\,. \tag{2.4.146}$$

Prove by adapting the method of types of Section 2.1.1 that $L_n^{\mathbf{X},(2)}$ satisfies a large deviations principle with the rate function $I^{(2)}(\cdot)$ specified in (2.4.137).

Remark: The estimates (2.4.145) and (2.4.146) are consequences of a somewhat involved combinatorial estimate of $|T(q)|$ (see for example [18], eq. (35)-(37) and references therein)

2.5 Long rare segments in random walks

Let $X_1,\ldots,X_n,\ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $\Lambda(\cdot)$ a steep function such that $\Lambda(\lambda)<\infty$ for some open ball around the origin. Let A be any rare Λ^* continuity subset of $\mathbb{R}^d$, namely, such that

$$I_A\stackrel{\Delta}{=}\inf_{x\in\bar{A}}\Lambda^*(x)=\inf_{x\in A^o}\Lambda^*(x)>0\,, \tag{2.5.147}$$

where $\Lambda^*(x)$ is the Legendre transform of $\Lambda(\lambda)$ defined in (2.3.60).

Consider the random walk $S_k=\sum_{i=1}^k X_i$, $k=1,2,\cdots$, $S_0=0$ and let

$$R_n^{(A)}\stackrel{\Delta}{=}\max\left\{m-k:\ 0\le k<m\le n\,,\ \ \frac{S_m-S_k}{m-k}\in A\right\}\,. \tag{2.5.148}$$

Thus, $R_n^{(A)}$ is the maximal length among all segments of the random walk up to n in which the empirical mean is within the set A.

Associated with $R_n^{(A)}$ is the dual variable

$$T_r^{(A)} \triangleq \inf\left\{m : \frac{S_m - S_k}{m-k} \in A \quad \text{for some} \quad 0 \le k \le m-r\right\}, \tag{2.5.149}$$

so that $\{R_n^{(A)} \ge r\}$ if and only if $\{T_r^{(A)} \le n\}$.

The analysis of the random variables $R_n^{(A)}$ and $T_r^{(A)}$ has applications in problems of the statistical analysis of DNA sequence matching and in the analysis of search algorithms in computer science. The following Theorem yields estimates on rare events which are usually associated with the probability of errors for matching algorithms. For some applications and refinements of these estimates, c.f. [23] and the exercises at the end of this section.

Theorem 2.5.1 $\lim_{n\to\infty}(R_n^{(A)}/\log n) = \lim_{r\to\infty}(r/\log T_r^{(A)}) = 1/I_A$, *almost surely.*

Proof: By the Borel-Cantelli lemma and utilizing the duality of events $\{R_n^{(A)} \ge r\} \equiv \{T_r^{(A)} \le n\}$ the theorem follows from the estimates

$$\sum_{r=1}^{\infty} \text{Prob}\,(T_r^{(A)} \le e^{r(I_A-\epsilon)}) < \infty, \qquad \forall \epsilon > 0 \tag{2.5.150}$$

$$\sum_{r=1}^{\infty} \text{Prob}\,(T_r^{(A)} > e^{r(I_A+\epsilon)}) < \infty, \qquad \forall \epsilon > 0 \tag{2.5.151}$$

when $I_A < \infty$ and from

$$\sum_{r=1}^{\infty} \text{Prob}\,(T_r^{(A)} \le e^{r/\epsilon}) < \infty, \qquad \forall \epsilon > 0 \tag{2.5.152}$$

when $I_A = \infty$.

The desired estimates (2.5.150), (2.5.151), and (2.5.152), are immediate consequences of the bounds

$$\text{Prob}(T_r^{(A)} > n) \le e^{-\lfloor \frac{n}{r} \rfloor \mu_r(A)} \tag{2.5.153}$$

and

$$\text{Prob}(T_r^{(A)} \le n) \le n \sum_{\ell=r}^{\infty} \mu_\ell(A), \tag{2.5.154}$$

coupled with Cramer's theorem (Theorem 2.3.2)

$$\lim_{\ell\to\infty}\frac{1}{\ell}\log\mu_\ell(A) = -I_A\,, \tag{2.5.155}$$

where μ_ℓ denotes the law of $\hat{S}_\ell = \frac{1}{\ell}S_\ell$ for $\ell \in Z^+$ and $\lfloor\frac{n}{r}\rfloor$ denotes the largest integer which is not larger than $\frac{n}{r}$. Indeed, assuming (2.5.154), one has by substituting $n = \lfloor e^{r(I_A-\epsilon)}\rfloor$ that when $I_A < \infty$

$$\sum_{r=1}^{\infty}\mathrm{Prob}\,(T_r^{(A)} \le e^{r(I_A-\epsilon)}) \le \sum_{r=1}^{\infty} e^{r(I_A-\epsilon)}\sum_{\ell=r}^{\infty} ce^{-\ell(I_A-\epsilon/2)} \le c'\sum_{r=1}^{\infty} e^{-r\epsilon/2} < \infty$$

for some positive constants c, c'. When $I_A = \infty$ one obtains (2.5.152) by choosing $n = \lfloor e^{r/\epsilon}\rfloor$ in (2.5.154) and following the same line of proof. Similarly, starting with (2.5.153) and choosing $n = \lfloor e^{r(I_A+\epsilon)}\rfloor$ one obtains

$$\sum_{r=1}^{\infty}\mathrm{Prob}\,(T_r^{(A)} > e^{r(I_A+\epsilon)}) \le \sum_{r=1}^{\infty}\exp\left(-\frac{c}{r}e^{r(I_A+\epsilon)}e^{-r(I_A+\epsilon/2)}\right) \le \sum_{r=1}^{\infty}\exp(-c''e^{c'r}) < \infty\,,$$

for some positive constants c, c', c''. We turn therefore to the proof of the bounds (2.5.153) and (2.5.154). The bound (2.5.153) follows by the inclusion

$$\{T_r^{(A)} \le n\} \supset \bigcup_{\ell=1}^{\lfloor\frac{n}{r}\rfloor} B_\ell\,, \qquad B_\ell \triangleq \left\{\frac{1}{r}(S_{\ell r} - S_{(\ell-1)r}) \in A\right\}\,, \tag{2.5.156}$$

as it implies

$$\mathrm{Prob}(T_r^{(A)} > n) \le 1 - \mathrm{Prob}\left(\bigcup_{\ell=1}^{\lfloor\frac{n}{r}\rfloor} B_\ell\right) = (1-\mathrm{Prob}(B_1))^{\lfloor\frac{n}{r}\rfloor} \le e^{-\lfloor\frac{n}{r}\rfloor\,\mathrm{Prob}(B_1)} \tag{2.5.157}$$

due to the fact that $\{B_\ell\}_{\ell=1}^{\infty}$ are independent events related to disjoint segments of the random walk, of equal probabilities $\mathrm{Prob}(B_\ell) = \mu_r(A), \ell = 1, 2, \cdots$.

The bound (2.5.154) follows by the inclusion

$$\{T_r^{(A)} \le n\} \subset \bigcup_{k=0}^{n-r}\bigcup_{m=k+r}^{n} C_{k,m} \subset \bigcup_{k=0}^{n-1}\bigcup_{m=k+r}^{\infty} C_{k,m}\,,\ C_{k,m} \triangleq \left\{\frac{S_m - S_k}{m-k} \in A\right\}\,, \tag{2.5.158}$$

and the union of events bound. Note that $\mathrm{Prob}\,(C_{k,m}) = \mu_{m-k}(A)$, and $m-k \ge r$ while in (2.5.158), there are at most n possible choices of k. □

Remark: Note that Theorem 2.5.1 holds as long as (2.5.155) holds. For example, consider exercises 2.5.3 and 2.5.4.

Exercises:

2.5.1 Suppose that $I_A < \infty$ and (2.5.155) may be refined to

$$\lim_{r\to\infty}[\mu_r(A)r^{d/2}e^{rI_A}] = a \ ,$$

for some $a \in (0,\infty)$ (such an example is presented in Section 2.10 for $d = 1$). Let

$$\hat{R}_n^{(A)} \triangleq \frac{I_A R_n^{(A)} - \log n}{\log\log n} + \frac{d}{2} \ .$$

(a). Repeat the above calculations and prove that $\limsup_{n\to\infty} |\hat{R}_n^{(A)}| \leq 1$ almost surely.
(b). Also deduce that $\lim_{n\to\infty} \text{Prob}(\hat{R}_n^{(A)} \leq -\epsilon) = 0$ for all $\epsilon > 0$.

2.5.2 Let $A = \{1\}$ and X_i be i.i.d. Bernoulli(p) random variables. Then, $R_n^{(A)}$ is the longest consecutive run of 1-s in the binary sequence $X_1,\ldots,X_n$. Let the renewal times $Z_1, Z_2, \cdots$ be the locations of zeros in this sequence (with $Z_0 \triangleq 0$). Then, $Q_k \triangleq Z_k - Z_{k-1} - 1$, $k = 1,2,\cdots$ are i.i.d. Geometric($1-p$) random variables and $R_{Z_k}^{(A)} = \max\{Q_1,\ldots,Q_k\}$. By standard renewal theory $\frac{Z_k}{k} \to \frac{1}{1-p}$ as $k \to \infty$, almost surely. Verify that here $I_A = -\log p$ and deduce that

$$\lim_{n\to\infty} \epsilon_n(I_A R_n^{(A)} - \log n) = 0$$

almost surely, whenever $\lim_{n\to\infty} \epsilon_n = 0$.

2.5.3 (a). Consider a sequence $X_1,\ldots,X_n$ of i.i.d. random variables over a finite alphabet Σ having a marginal law μ such that $\Sigma_\mu = \Sigma$. Let $R_n^{(\Gamma)}$ be the longest among all segments of this sequence with segmental empirical measures in the *open* set $\Gamma \subset M_1(\Sigma)$. Assume that $\mu \notin \overline{\Gamma}$ and derive the analog of Theorem 2.5.1 for this situation.
(b). Assume further that Γ is convex. Let ν^* be the unique minimizer of $H(\cdot|\mu)$ in $\overline{\Gamma}$. Prove that as $n \to \infty$ the empirical measures associated with the segments contributing to $R_n^{(\Gamma)}$ converge almost surely to ν^*.
Hint: Let $\Gamma_\delta \triangleq \Gamma \cap \overline{B}^c_{\nu^*,\delta}$ where $B_{\nu^*,\delta}$ is an open ball of radius $\delta > 0$ around ν^*. Prove that $\limsup_{n\to\infty} \frac{R_n^{(\Gamma_\delta)}}{R_n^{(\Gamma)}} < 1$ almost surely, for any $\delta > 0$. Deduce that for any $\delta > 0$ the empirical measures of the segments contributing to $R_n^{(\Gamma)}$ are eventually within distance δ of ν^* almost surely.

2.5.4 Assume that $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$ are as in exercise 2.4.1. Specifically, X_k are the states of a Markov chain over the finite set $\{1,2,\ldots,|\Sigma|\}$ with an irreducible transition matrix Π,

and the conditional law of Y_k when $X_k = j$ is $\mu_j \in M_1(\mathbb{R}^d)$ while $\{Y_k\}$ are independent given any realization of the Markov chain states. Further, suppose that the logarithmic moment generating functions Λ_j associated with μ_j are finite everywhere and define the matrices Π_λ via

$$\pi_\lambda(i,j) \stackrel{\Delta}{=} \pi(i,j)\, e^{\Lambda_j(\lambda)} \,.$$

Let $\Lambda^*(x)$ denote the Legendre transform of $\log \rho(\Pi_\lambda)$ and suppose A is a rare Λ^* continuity set. Let $R_n^{(A)}$ be the longest among all segments whose Y-segmental empirical mean belongs to $A \subset \mathbb{R}^d$. Prove that Theorem 2.5.1 holds with Λ^* as defined here.

2.6 The Gibbs conditioning principle in finite alphabet

Let $X_1, X_2, \ldots X_n$ be a sequence of i.i.d. random variables with law μ over the finite alphabet $\Sigma \subset \mathbb{R}^1$ and assume without loss of generality that $\Sigma_\mu = \Sigma$. The following question is of fundamental importance in statistical mechanics. Given a set $A \in \mathbb{R}$ and a constraint of the type $\hat{S}_n \in A$ what is the conditional law of X_1 for large n ? In other words, what are the limit points of the conditional probability vector

$$\mu_n^*(a_i) \stackrel{\Delta}{=} \text{Prob}_\mu(X_1 = a_i \,|\, \hat{S}_n \in A) \;\; i = 1, \ldots, |\Sigma| \,, \tag{2.6.159}$$

as $n \to \infty$ (recall that $\hat{S}_n \stackrel{\Delta}{=} \frac{1}{n}\sum_{j=1}^n X_j = < L_n^{\mathbf{X}}, \mathrm{a} >$). Note that for any function $f : \Sigma \to \mathbb{R}^1$

$$\begin{aligned} < \mu_n^*, f > &= E[f(X_1) \mid \hat{S}_n \in A] = E[f(X_2) \mid \hat{S}_n \in A] \\ &= E[\frac{1}{n}\sum_{j=1}^n f(X_j) \mid \hat{S}_n \in A] = E[< L_n^{\mathbf{X}}, f > \;\mid\; < L_n^{\mathbf{X}}, \mathrm{a} > \in A] \end{aligned} \tag{2.6.160}$$

where we have used the fact that X_j are identically distributed (although not independent) even under the conditioning $\hat{S}_n \in A$. Therefore,

$$\mu_n^* = E[L_n^{\mathbf{X}} \mid L_n^{\mathbf{X}} \in \Gamma] \,, \tag{2.6.161}$$

where $\Gamma \stackrel{\Delta}{=} \{\nu :< \nu, \mathrm{a} > \in A\}$ (compare with (2.1.19) in Section 2.1.2). Therefore, the characterization of possible limit points of the sequence μ_n^* as $n \to \infty$ can be cast in terms of conditional limits for the empirical measures $L_n^{\mathbf{X}}$.

The following characterization of the limits of μ_n^* is a consequence of Theorem 2.1.1 for any non-empty set Γ which is an $H(\cdot|\mu)$ continuity set, namely,

$$I_\Gamma \stackrel{\Delta}{=} \inf_{\nu\in\Gamma^o} H(\nu|\mu) = \inf_{\nu\in\overline{\Gamma}} H(\nu|\mu)\,. \tag{2.6.162}$$

Theorem 2.6.1 (Gibb's principle)

(a). The set of possible limit points of μ_n^* is the closure of the convex hull of

$$\mathcal{M}_\Gamma \stackrel{\Delta}{=} \{\nu\in\overline{\Gamma} : H(\nu|\mu) = I_\Gamma\} \tag{2.6.163}$$

(b). For any convex set Γ of non-empty interior the set $\mathcal{M}_\Gamma$ is a point to which μ_n^* converges as $n\to\infty$.

Remark: For conditions on Γ (alternatively, on A) under which (2.6.162) holds see exercises 2.1.1-2.1.3 in Section 2.1.1.

Proof: As $|\Sigma| < \infty$, $\overline{\Gamma}$ is a compact set and thus $\mathcal{M}_\Gamma$ is non-empty. Moreover, part (b) of the theorem follows from part (a) by exercise 2.1.3 and the compactness of $M_1(\Sigma)$ (in that exercise you showed that indeed (2.6.162) holds when Γ is a convex set of non-empty interior and that the set $\mathcal{M}_\Gamma$ is a point).

We shall prove that for any $\delta > 0$

$$\lim_{n\to\infty} \text{Prob}(L_n^{\mathbf{X}} \in \mathcal{M}_\Gamma^\delta \,|\, L_n^{\mathbf{X}} \in \Gamma) = 1\,, \tag{2.6.164}$$

with an exponential (in n) rate of convergence, where $\mathcal{M}_\Gamma^\delta \stackrel{\Delta}{=} \{\nu : d_V(\nu,\mathcal{M}_\Gamma) < \delta\}$.

Since $M_1(\Sigma)$ is a bounded set, (2.6.161) and (2.6.164) imply that for any $\delta > 0$, μ_n^* eventually belongs to the convex hull of $\mathcal{M}_\Gamma^{2\delta}$. All points in the convex hull of $\mathcal{M}_\Gamma^\delta$ are within variational distance δ of some point in the convex hull of $\mathcal{M}_\Gamma$ (since d_V is a convex function on $M_1(\Sigma)\times M_1(\Sigma)$). Thus, since δ is arbitrarily small, limit points of μ_n^* are necessarily in the closure of the convex hull of $\mathcal{M}_\Gamma$ as claimed.

The limit (2.6.164) definitely follows from

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\log \text{Prob}_\mu(L_n^{\mathbf{X}} \in (\mathcal{M}_\Gamma^\delta)^c \,|\, L_n^{\mathbf{X}} \in \Gamma) \;&=\\ \limsup_{n\to\infty} \left\{\frac{1}{n}\log \text{Prob}_\mu(L_n^{\mathbf{X}} \in (\mathcal{M}_\Gamma^\delta)^c \cap \Gamma) \;-\; \frac{1}{n}\log \text{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma)\right\} &< 0\,.\end{aligned} \tag{2.6.165}$$

However, by Theorem 2.1.1 and (2.6.162)

$$I_\Gamma = -\lim_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in \Gamma) \ , \tag{2.6.166}$$

whereas by Theorem 2.1.1 also

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{Prob}_\mu(L_n^{\mathbf{X}} \in (\mathcal{M}_\Gamma^\delta)^c \cap \Gamma) \leq -\inf_{\nu\in(\mathcal{M}_\Gamma^\delta)^c\cap\overline{\Gamma}} H(\nu|\mu) \ . \tag{2.6.167}$$

Observe that $\mathcal{M}_\Gamma^\delta$ are open sets and therefore $(\mathcal{M}_\Gamma^\delta)^c \cap \overline{\Gamma}$ are compact sets. Thus, for some $\tilde{\nu} \in (\mathcal{M}_\Gamma^\delta)^c \cap \overline{\Gamma}$

$$\inf_{\nu\in(\mathcal{M}_\Gamma^\delta)^c\cap\overline{\Gamma}} H(\nu|\mu) \ = \ H(\tilde{\nu}\,|\,\mu) > I_\Gamma \ , \tag{2.6.168}$$

where the above inequality follows from the definition of $\mathcal{M}_\Gamma$ since $\tilde{\nu} \notin \mathcal{M}_\Gamma$ while $\tilde{\nu} \in \overline{\Gamma}$. Finally, (2.6.164) follows from (2.6.165)-(2.6.168). □

Remarks

(a). Intuitively one expects $X_1, \ldots, X_k$ to be asymptotically independent (as $n \to \infty$) for any fixed k, when the conditioning event is $\{L_n^{\mathbf{X}} \in \Gamma\}$. This is indeed shown in exercise 2.6.3 by considering "super-symbols" from the enlarged alphabet Σ^k.

(b). Theorem 2.6.1 holds for any set Γ satisfying (2.6.162). However, the particular conditioning set $\{\nu : < \nu, \mathbf{a} > \in A\}$ has an important significance in statistical mechanics because it represents an energy-like constraint.

(c). Recall the relationship (2.1.23) of Section 2.1.2 which implies that for any non-empty, convex, open set $A \subset K$ the unique limit of μ_n^* is of the form

$$\nu_\lambda(a_i) = e^{\lambda a_i - \Lambda(\lambda)} \mu(a_i) \ ,$$

for some appropriately chosen $\lambda \in \mathbb{R}^1$ which is called the **Gibbs parameter** associated with A. In particular, for any $x \in K^o$ the Gibbs parameter λ associated with $A = (x - \delta, x + \delta)$ converges as $\delta \to 0$ to $\lambda^{(x)}$, the unique solution of the equation $\Lambda'(\lambda) = x$ (for details see Section 2.1.2).

(d). A Gibbs conditioning principle holds beyond the i.i.d. case. Actually, all that is needed is that X_i are exchangeable conditionally upon any given value of $L_n^{\mathbf{X}}$ (so that (2.6.161) holds). For such an example, consider exercise 2.6.2.

Exercises:

2.6.1 Prove Theorem 2.6.1 by the method of types (specifically, use Lemma 2.1.4 directly).

2.6.2 Prove the Gibbs conditioning principle for sampling without replacement.
(a). Observe that again X_j are identically distributed even when $L_n^{\mathbf{X}}$ is given. Conclude that (2.6.161) holds.
(b). Assume that Γ is such that

$$I_\Gamma = \inf_{\nu \in \Gamma^o} I_{\beta,\mu}(\nu) = \inf_{\nu \in \overline{\Gamma}} I_{\beta,\mu}(\nu) < \infty \, ,$$

and define $\mathcal{M}_\Gamma \stackrel{\Delta}{=} \{\nu \in \overline{\Gamma} : I_{\beta,\mu}(\nu) = I_\Gamma\}$. Prove that now both parts of Theorem 2.6.1 hold (for part (b) you may relay on exercise 2.1.12).

2.6.3 (a). Suppose that $\Sigma = (\Sigma')^k$ and $\mu = (\mu')^k$ are a k-th product alphabet and a k-th product underlying measure on it and assume that $\Sigma'_{\mu'} = \Sigma'$ (as usual). For any law $\nu \in M_1(\Sigma)$ let $\nu^{(j)} \in M_1(\Sigma')$, $j = 1, \ldots, k$ denote its j-th marginal on Σ'. Prove that

$$\frac{1}{k} H(\nu \,|\, \mu) \geq \frac{1}{k} \sum_{j=1}^{k} H(\nu^{(j)} \,|\, \mu') \geq H(\frac{1}{k} \sum_{j=1}^{k} \nu^{(j)} \,|\, \mu') \, ,$$

with equality if and only if $\nu = (\nu')^k$ for some $\nu' \in M_1(\Sigma')$.
(b). Assume that

$$\Gamma \stackrel{\Delta}{=} \{\nu : \frac{1}{k} \sum_{j=1}^{k} \nu^{(j)} \in \Gamma'\} \tag{2.6.169}$$

for some $\Gamma' \subset M_1(\Sigma')$ which satisfies (2.6.162) with respect to μ'. Prove that then $\mathcal{M}_\Gamma = (\mathcal{M}_{\Gamma'})^k$ and conclude that any limit point of μ_n^* is a k-th product of some appropriate law on Σ'.
(c). Consider now the k-th joint conditional law

$$\mu_n^*(a'_{i_1}, \ldots, a'_{i_k}) \stackrel{\Delta}{=} \mathrm{Prob}_{\mu'}(X_1 = a'_{i_1}, \ldots, X_k = a'_{i_k} \,|\, L_n^{\mathbf{X}} \in \Gamma') \;\; a'_{i_j} \in \Sigma' \, , \; j = 1, \ldots, k \, ,$$

where X_i are i.i.d. with marginal law $\mu' \in M_1(\Sigma')$ over the finite alphabet Σ' and $\Gamma' \subset M_1(\Sigma')$ satisfies (2.6.162). Let $\mu = (\mu')^k$ be the law of $Y_i = (X_{1+k(i-1)}, \ldots, X_{ki})$ over a new alphabet Σ. Prove that for any $m \in Z^+$

$$\mu_{mk}^*(a_{\mathbf{i}}) = \mathrm{Prob}_\mu(Y_1 = a_{\mathbf{i}} \,|\, L_m^{\mathbf{Y}} \in \Gamma) \, , \;\; \forall \, a_{\mathbf{i}} \in \Sigma \, ,$$

where Γ is defined in (2.6.169). Deduce that as $n \to \infty$ along integer multiples of k the random variables X_i, $i = 1, \ldots, k$ are asymptotically conditionally i.i.d (i.e., any limit point of μ_n^* is a k-th product of an element of $M_1(\Sigma')$).

(d). Prove that the above conclusion extends to n which need not be an integer multiple of k whenever $\mathcal{M}_{\Gamma'}$ is a single point.

2.7 The hypothesis test problem

Consider the problem of hypothesis testing between two product measures for the i.i.d. random variables $Y_1, \cdots, Y_n, \cdots$. Specifically, Y_j are either distributed according to the law $\mu_0 \in M_1(\Sigma)$ (hypothesis H_0) or according to $\mu_1 \in M_1(\Sigma)$ (hypothesis H_1). The alphabet Σ may in general be quite arbitrary provided that the probability measures μ_0 and μ_1 are well defined (Markov chains over finite alphabet are considered in exercise 2.7.4).

Definition 2.7.1 *A decision test $\mathcal{S}$ is a sequence of maps $\mathcal{S}^n : \Sigma^n \to \{0,1\}$, for $n = 1, 2, \cdots$, with the interpretation that when $Y_1 = y_1, \ldots, Y_n = y_n$ is observed then H_0 is accepted (H_1 rejected) if $\mathcal{S}^n(y_1, \ldots, y_n) = 0$ while H_1 is accepted (H_0 rejected) if $\mathcal{S}^n(y_1, \ldots, y_n) = 1$.*

The performance of a decision test $\mathcal{S}$ is determined by the error probabilities

$$\alpha_n \triangleq \text{Prob}_{\mu_0}(H_0 \text{ rejected by } \mathcal{S}^n)\,, \quad \beta_n \triangleq \text{Prob}_{\mu_1}(H_1 \text{ rejected by } \mathcal{S}^n)\,, \quad n \in Z^+\,. \tag{2.7.170}$$

One wishes to minimize β_n. If no constraint is put on α_n, then one may have $\beta_n = 0$ with the test $\mathcal{S}^n(y_1, \ldots, y_n) \equiv 1$ at the cost of $\alpha_n = 1$. Thus, a sensible criterion for optimality is to seek a test which minimizes β_n subject to a constraint on α_n. Suppose now that the probability measures μ_0, μ_1 are known a-priori and that they are *equivalent measures*, so the likelihood ratios $L_{0||1}(y) = \frac{d\mu_0}{d\mu_1}(y)$ and $L_{1||0}(y) = \frac{d\mu_1}{d\mu_0}(y)$ exist (some extensions for μ_0, μ_1 which are not equivalent are given in exercise 2.7.3). This equivalence assumption is valid for example when μ_0, μ_1 are discrete measures with $\Sigma_{\mu_0} = \Sigma_{\mu_1}$, or when $\Sigma = \mathbb{R}^d$ and both μ_0 and μ_1 possess strictly positive densities. In order to avoid trivialities it is further assumed that μ_0 and μ_1 are distinguishable, i.e. that they differ on a set whose probability under μ_0 (and μ_1) is positive.

Let $X_j \triangleq \log L_{1||0}(Y_j) = -\log L_{0||1}(Y_j)$ denote the observed log-likelihood ratios. These are bona-fide i.i.d. random variables over $\mathbb{R}^1$ which are non-zero with positive probability. Moreover,

$$\overline{x}_0 \triangleq E_{\mu_0}[X_1] = E_{\mu_1}[X_1 e^{-X_1}] ,$$

exists (with possibly $\overline{x}_0 = -\infty$) as $xe^{-x} \leq 1$. Similarly,

$$\overline{x}_1 \triangleq E_{\mu_1}[X_1] = E_{\mu_0}[X_1 e^{X_1}] > E_{\mu_0}[X_1] = \overline{x}_0 ,$$

exists (with possibly $\overline{x}_1 = \infty$) and the above inequality is strict since X_1 is non-zero with positive probability. Note that $\overline{x}_0$ and $\overline{x}_1$ may be both characterized in terms of relative entropy, c.f. exercise 2.7.1.

Definition 2.7.2 *A Neyman-Pearson test is a test in which for any n the mean observed log-likelihood ratio $\hat{S}_n \triangleq \frac{1}{n}\sum_{j=1}^n X_j$ is compared against a threshold γ_n and H_1 is accepted (rejected) when $\hat{S}_n > \gamma_n$ ($\hat{S}_n \leq \gamma_n$).*

It is well known that Neyman-Pearson tests are optimal in the sense that there are neither tests with the same value of α_n and a smaller value of β_n nor tests with the same value of β_n and a smaller value of α_n (see for example [7] for a simple proof of this claim).

The exponential rates of α_n and β_n for Neyman-Pearson tests with constant thresholds $\gamma \in (\overline{x}_0, \overline{x}_1)$ are thus of particular interest. These may be cast in terms of the large deviations of $\hat{S}_n$. In particular, since X_j are i.i.d. real valued random variables, the following theorem is an application of Theorem 2.2.1.

Theorem 2.7.1 *For any Neyman-Pearson test with constant threshold $\gamma \in (\overline{x}_0, \overline{x}_1)$*

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = -\Lambda_0^*(\gamma) < 0 , \tag{2.7.171}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = \gamma - \Lambda_0^*(\gamma) < 0 , \tag{2.7.172}$$

where

$$\Lambda_0^*(x) \triangleq \sup_{\lambda\in[0,1]} \{\lambda x - \Lambda_0(\lambda)\} \tag{2.7.173}$$

and $\Lambda_0(\lambda) = \log E_{\mu_0}[e^{\lambda X_1}]$ is the logarithmic moment generating function of X_1 under H_0.

Proof: Both (2.7.171) and (2.7.172) follow by slight modification of the proof of Theorem 2.2.1. First note that

$$\overline{x}_0 = \lim_{\lambda \downarrow 0} \Lambda_0'(\lambda) < \gamma < \lim_{\lambda \uparrow 1} \Lambda_0'(\lambda) = \overline{x}_1 \ .$$

Thus, $\gamma = \Lambda_0'(\eta)$ for some $\eta \in (0,1)$ and $\Lambda_0^*(\gamma)$ indeed equals the Legendre transform of Λ_0 at the point γ. Now by (2.2.46)

$$\limsup_{n \to \infty} \frac{1}{n} \log \alpha_n = \limsup_{n \to \infty} \frac{1}{n} \log \text{Prob}_{\mu_0}(\hat{S}_n \in (\gamma, \infty)) \leq - \inf_{x \geq \gamma} \Lambda_0^*(x) = -\Lambda_0^*(\gamma) \qquad (2.7.174)$$

where the last equality follows since $\Lambda_0^*(\cdot)$ is nondecreasing on $[\gamma, \infty)$ as $\gamma > \overline{x}_0$.

By the definition of X_j the logarithmic moment generating function associated with μ_1 is merely $\Lambda_0(\lambda + 1)$ and so the rate function $\Lambda_1^*(x) \stackrel{\Delta}{=} \Lambda_0^*(x) - x$ governs the large deviations bounds for the laws of $\hat{S}_n$ under H_1. Apply (2.2.46) once again to obtain

$$\limsup_{n \to \infty} \frac{1}{n} \log \beta_n = \limsup_{n \to \infty} \frac{1}{n} \log \text{Prob}_{\mu_1}(\hat{S}_n \in (-\infty, \gamma]) \leq - \inf_{x \leq \gamma} \Lambda_1^*(x) = -\Lambda_1^*(\gamma) \qquad (2.7.175)$$

where the last equality follows since $\Lambda_1^*(\cdot)$ is nonincreasing on $(-\infty, \gamma]$ as $\gamma < \overline{x}_1$.

Since $\gamma = \Lambda_0'(\eta)$ for some $\eta \in (0,1)$ where Λ_0 is a strictly convex, C^∞ function, both $\Lambda_0^*(\cdot)$ and $\Lambda_1^*(\cdot)$ are continuous at the point γ (consider further exercise 2.2.5 of Section 2.2). Moreover, for large enough r the lower bound (2.2.51) applies to $y_r = \gamma + \frac{1}{r}$ and $\delta_r = \frac{1}{r}$ implying that

$$\liminf_{n \to \infty} \frac{1}{n} \log \alpha_n \geq \liminf_{n \to \infty} \frac{1}{n} \log \text{Prob}_{\mu_0}(\hat{S}_n \in B_{y_r, \delta_r}) \geq -\Lambda_0^*(y_r) \qquad (2.7.176)$$

where $B_{y,\delta} \stackrel{\Delta}{=} (y - \delta, y + \delta)$. By taking the limit $r \to \infty$ and combining the above lower bound with (2.7.174) one deduces (2.7.171). Similarly, one has for $z_r = \gamma - \frac{1}{r}$ (with $\delta_r = \frac{1}{r}$ and r large)

$$\liminf_{n \to \infty} \frac{1}{n} \log \beta_n \geq \liminf_{n \to \infty} \frac{1}{n} \log \text{Prob}_{\mu_1}(\hat{S}_n \in B_{z_r, \delta_r}) \geq -\Lambda_1^*(z_r) \qquad (2.7.177)$$

implying (2.7.172) in the limit $r \to \infty$. □

Remarks: (a). Observe that Theorem 2.7.1 holds even when $\overline{x}_0 = -\infty$ or $\overline{x}_1 = \infty$ or both. Its proof is actually a specialization of exercise 2.2.7 from Section 2.2.

(b). A refinement of Theorem 2.7.1 is given in exercise 2.10.3 where the exact limiting behavior of α_n (β_n) is derived.

A corollary of Theorem 2.7.1 is Chernoff's asymptotic bound on the best achievable Bayesian probability of error

$$P_n^{(e)} = \text{Prob}(H_0)\alpha_n + \text{Prob}(H_1)\beta_n \qquad (2.7.178)$$

Corollary 2.7.1 (Chernoff's Bound) *If* $0 < \text{Prob}(H_0) < 1$ *then*

$$\liminf_{n\to\infty}\{\frac{1}{n}\log P_n^{(e)}\} = -\Lambda_0^*(0) = -\inf_{\lambda\in[0,1]}\Lambda_0(\lambda) \quad , \tag{2.7.179}$$

where the above infimum is over all tests.

Remarks:

(a). In particular, Theorem 2.7.1 thus implies that the best Bayesian exponential error rate is achieved by a Neyman-Pearson test with *zero threshold.*

(b). The rate $\Lambda_0^*(0)$ is called Chernoff's information of the measures μ_0 and μ_1.

Proof: It suffices to consider only Neyman-Pearson tests. Let α_n^* and β_n^* be the error probabilities for the zero threshold Neyman-Pearson test. Then by (2.7.171) and (2.7.172)

$$\lim_{n\to\infty}\frac{1}{n}\log\alpha_n^* = \lim_{n\to\infty}\frac{1}{n}\log\beta_n^* = -\Lambda_0^*(0) \tag{2.7.180}$$

For any test either $\alpha_n \geq \alpha_n^*$ (when $\gamma_n \leq 0$), or $\beta_n \geq \beta_n^*$ (when $\gamma_n \geq 0$). Thus, for any test

$$\frac{1}{n}\log P_n^{(e)} \geq \frac{1}{n}\log[\min\{\text{Prob}(H_0)\,,\,\text{Prob}(H_1)\}] + \min\{\frac{1}{n}\log\alpha_n^*\,,\,\frac{1}{n}\log\beta_n^*\}\,.$$

As $0 < \text{Prob}(H_0) < 1$, the limit $n \to \infty$ yields (2.7.179) in view of (2.7.180). □

Another corollary of Theorem 2.7.1 is the following lemma which determines the best exponential rate for β_n when α_n are bounded away from 1.

Lemma 2.7.1 (Stein's Lemma) *Let* β_n^ϵ *be the minimum of* β_n *among all tests with* $\alpha_n < \epsilon$. *Then, for any* $\epsilon < 1$

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n^\epsilon = \overline{x}_0 \quad . \tag{2.7.181}$$

Proof: It clearly suffices to consider only Neyman-Pearson tests. Then,

$$\alpha_n = \text{Prob}_{\mu_0}(\hat{S}_n > \gamma_n) \tag{2.7.182}$$

and

$$\beta_n = \text{Prob}_{\mu_1}(\hat{S}_n \leq \gamma_n) = E_{\mu_1}[1_{\hat{S}_n\leq\gamma_n}] = E_{\mu_0}[1_{\hat{S}_n\leq\gamma_n}e^{n\hat{S}_n}]\,, \tag{2.7.183}$$

where the last equality follows from the definition of X_j (as the observed log-likelihood ratios). The identity (2.7.183) yields

$$\frac{1}{n}\log\beta_n = \frac{1}{n}\log E_{\mu_0}[1_{\hat{S}_n\leq\gamma_n}e^{n\hat{S}_n}] \leq \gamma_n \tag{2.7.184}$$

(a). Suppose first that $\overline{x}_0 = -\infty$. Then, for any Neyman-Pearson test with a fixed threshold γ, eventually $\alpha_n < \epsilon$ by (2.7.171). Thus, $\frac{1}{n} \log \beta_n^\epsilon \leq \gamma$ for any γ and n large enough by (2.7.184) and (2.7.181) follows.

(b). Assume now that $\overline{x}_0 > -\infty$. Similarly, apply (2.7.171) to deduce that eventually $\alpha_n < \epsilon$ for Neyman-Pearson tests with a constant threshold $\gamma > \overline{x}_0$ and so by (2.7.184)

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon \leq \overline{x}_0 + \eta \,, \tag{2.7.185}$$

for any $\eta > 0$ and any $\epsilon > 0$.

Moreover, without loss of generality one may assume that

$$\liminf_{n\to\infty} \gamma_n \geq \overline{x}_0 \,, \tag{2.7.186}$$

for otherwise, by the weak law of large numbers $\limsup_{n\to\infty} \alpha_n = 1$. When (2.7.186) holds and $\alpha_n < \epsilon$ then by (2.7.182) and the weak law of large numbers

$$\liminf_{n\to\infty} \text{Prob}_{\mu_0}(\hat{S}_n \in [\overline{x}_0 - \eta \,, \gamma_n]) \geq 1 - \epsilon \quad \text{for any} \quad \eta > 0 \,. \tag{2.7.187}$$

Hence, by (2.7.183)

$$\begin{aligned} \frac{1}{n} \log \beta_n &\geq \frac{1}{n} \log E_{\mu_0}[1_{\hat{S}_n \in [\overline{x}_0 - \eta \,, \gamma_n]} e^{n\hat{S}_n}] \\ &\geq \overline{x}_0 - \eta + \frac{1}{n} \log \text{Prob}_{\mu_0}(\hat{S}_n \in [\overline{x}_0 - \eta \,, \gamma_n]) \,. \end{aligned} \tag{2.7.188}$$

By combining (2.7.187), (2.7.188) and the optimality of these Neyman-Pearson tests one obtains

$$\liminf_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon \geq \overline{x}_0 - \eta \quad \text{for any} \quad \eta > 0 \,. \tag{2.7.189}$$

The desired limit (2.7.181) is now a direct consequence of (2.7.185) and (2.7.189). □

Exercises:

2.7.1 Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables over the finite set $\Sigma \triangleq \{a_1, \ldots, a_{|\Sigma|}\}$ and $\Sigma_{\mu_0} = \Sigma_{\mu_1} = \Sigma$ (namely, μ_0 and μ_1 are strictly positive over Σ).

(a). Prove that $\overline{x}_1 = H(\mu_1|\mu_0) < \infty$ and $\overline{x}_0 = -H(\mu_0|\mu_1) > -\infty$ (see Section 2.1 for the definition

of relative entropy).

(b). For $\eta \in [0,1]$ define the probability measures

$$\mu_\eta(a_j) \stackrel{\Delta}{=} \frac{\mu_1(a_j)^\eta \mu_0(a_j)^{1-\eta}}{\sum_{k=1}^{|\Sigma|} \mu_1(a_k)^\eta \mu_0(a_k)^{1-\eta}} \quad j = 1, \ldots, |\Sigma| \, .$$

Let $\gamma = H(\mu_\eta|\mu_0) - H(\mu_\eta|\mu_1)$ and prove that $\Lambda_0^*(\gamma) = H(\mu_\eta|\mu_0)$.

2.7.2 Consider the scenario of exercise 2.7.1.

(a). Define the conditional probability vectors

$$\mu_n^*(a_j) \stackrel{\Delta}{=} \mathrm{Prob}_{\mu_0}(Y_1 = a_j \,|\, H_0 \quad \text{rejected by} \quad \mathcal{S}^n) \;\; j = 1, \ldots, |\Sigma| \, , \tag{2.7.190}$$

where $\mathcal{S}$ is a Neyman-Pearson test with a fixed threshold $\gamma \stackrel{\Delta}{=} H(\mu_\eta|\mu_0) - H(\mu_\eta|\mu_1)$ and $\eta \in (0,1)$. Use Theorem 2.6.1 to deduce that $\mu_n^* \to \mu_\eta$ as $n \to \infty$.

Hint: You may also find parts of the proof of Theorem 2.1.2 useful for solving this problem.

(b). Consider now the k-th joint conditional law

$$\mu_n^*(a_{j_1}, \ldots, a_{j_k}) \stackrel{\Delta}{=} \mathrm{Prob}_{\mu_0}(Y_1 = a_{j_1}, \ldots, Y_k = a_{j_k} \,|\, H_0 \quad \text{rejected by} \quad \mathcal{S}^n) \;\; a_{j_\ell} \in \Sigma \, , \; \ell = 1, \ldots, k \, .$$

Apply exercise 2.6.3 in order to deduce that

$$\lim_{n \to \infty} \mu_n^*(a_{j_1}, \ldots, a_{j_k}) = \mu_\eta(a_{j_1}) \mu_\eta(a_{j_2}) \cdots \mu_\eta(a_{j_k}) \quad .$$

Try to interpret this result.

2.7.3 Suppose that $L_{1||0}(y) = \frac{d\mu_1}{d\mu_0}(y)$ does not exist while $L_{0||1}(y) = \frac{d\mu_0}{d\mu_1}(y)$ does exists. Prove that Stein's lemma holds true whenever $\overline{x}_0 \stackrel{\Delta}{=} -E_{\mu_0}[\log L_{0||1}(Y_1)] > -\infty$.

Hint: Split μ_1 into its singular part with respect to μ_0 and its restriction on the support of the measure μ_0.

2.7.4 Suppose that $Y_1, \ldots, Y_n$ are the states of a Markov chain over the finite set $\Sigma = \{1, 2, \ldots, |\Sigma|\}$ where the initial state of the chain Y_0 is known a-priori to be some $x \in \Sigma$. The transition matrix under H_0 is Π_0 while under H_1 it is Π_1, both of which are irreducible matrices with the same set of non-zero values. Here the Neyman-Pearson tests are based upon $X_j \stackrel{\Delta}{=} \log \frac{\pi_1(Y_{j-1}, Y_j)}{\pi_0(Y_{j-1}, Y_j)}$ and $\overline{x}_i = E_{\mu_i}[X_1]$ for $i = 0, 1$. Derive the analogs of Theorem 2.7.1 and Lemma 2.7.1 by using the results of Section 2.4.3.

2.8 Generalized maximum likelihood for finite alphabets

This section is devoted to yet another version of the hypothesis test problem presented in Section 2.7. In particular, the concept of decision test is as in definition 2.7.1 and the associated error probabilities are as given in (2.7.170) there. While the law μ_0 is again assumed known a-priori, here μ_1, the law of Y_j under the hypothesis H_1 is *unknown*. For that reason, neither the methods nor the results of Section 2.7 apply. Moreoever, one has to modify the error criterion since requiring uniformly small β_n over a possibly large class of μ_1 measures may be too strong (i.e., it may well be that no test can satisfy such a condition). It is reasonable therefore to ask for a criterion which involves asymptotic limits. For finite alphabets $\Sigma = \{a_1, \ldots, a_{|\Sigma|}\}$ such a criterion was suggested by Hoeffding, as follows.

Definition 2.8.1 *A test $\mathcal{S}$ is optimal (for a given $\eta > 0$) if, among all tests which satisfy*

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n \leq -\eta \tag{2.8.191}$$

the test $\mathcal{S}$ has maximal exponential rate of error, i.e. $-\limsup_{n\to\infty}\{\frac{1}{n}\log\beta_n\}$ is maximal (uniformly over all possible laws μ_1).

As will become evident in Section ?? a considerable weakening of this criterion is necessary for more general alphabets.

The following lemma states that it suffices to consider functions of the empirical measure when trying to construct an optimal test (i.e., the empirical measure is a sufficient statistic for this problem).

Lemma 2.8.1 *For any test $\mathcal{S}$ with error probabilities $\{\alpha_n, \beta_n\}_{n=1}^{\infty}$ there exists a test $\tilde{\mathcal{S}}$ with maps of the form $\tilde{\mathcal{S}}^n(\mathbf{x}) = \tilde{\mathcal{S}}(L_n^{\mathbf{x}}, n)$ whose error probabilities $\{\tilde{\alpha}_n, \tilde{\beta}_n\}_{n=1}^{\infty}$ satisfy*

$$\limsup_{n\to\infty} \frac{1}{n} \log \tilde{\alpha}_n \leq \limsup_{n\to\infty} \frac{1}{n} \log \alpha_n$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \tilde{\beta}_n \leq \limsup_{n\to\infty} \frac{1}{n} \log \beta_n \tag{2.8.192}$$

Proof: For any $n \in Z^+$ let $\mathcal{S}_0^n \triangleq (\mathcal{S}^n)^{-1}(0)$ and $\mathcal{S}_1^n \triangleq (\mathcal{S}^n)^{-1}(1)$ denote the subsets of Σ^n which the maps $\mathcal{S}^n$ assign to H_0 and H_1 respectively. For $i = 0, 1$ and any $\nu \in \mathcal{L}_n$ let $\mathcal{S}_i^{\nu,n} \triangleq \mathcal{S}_i^n \cap T(\nu)$

(recall that $T(\nu)$ is the type class of ν, see definition 2.1.2). Define

$$\tilde{S}(\nu,n) \triangleq \begin{cases} 0 & \text{if } |S_0^{\nu,n}| \geq \frac{1}{2}|T(\nu)| \\ 1 & \text{otherwise} \end{cases} \tag{2.8.193}$$

where $|A|$ denotes the cardinality of the set A. The test $\tilde{S}$ specified in the lemma's statement is composed of the maps $\tilde{S}^n(\mathbf{x}) = \tilde{S}(L_n^{\mathbf{x}}, n)$.

Recall that for any i.i.d. variables $\mathbf{X} = (X_1, \ldots, X_n)$ with marginal $\mu \in M_1(\Sigma)$ and for any possible type $\nu \in \mathcal{L}_n$, the conditional measure $\text{Prob}_\mu(\mathbf{X} \mid L_n^{\mathbf{X}} = \nu)$ is a uniform measure over the type class $T(\nu)$. In particular, if $\tilde{S}(\nu,n) = 0$ then

$$\frac{1}{2}\text{Prob}_{\mu_1}(L_n^{\mathbf{X}} = \nu) \leq \frac{|S_0^{\nu,n}|}{|T(\nu)|}\text{Prob}_{\mu_1}(L_n^{\mathbf{X}} = \nu) = \text{Prob}_{\mu_1}(\mathbf{X} \in S_0^{\nu,n}) \,. \tag{2.8.194}$$

Therefore

$$\begin{aligned} \tilde{\beta}_n &= \sum_{\{\nu:\tilde{S}(\nu,n)=0\}\cap\mathcal{L}_n} \text{Prob}_{\mu_1}(L_n^{\mathbf{X}} = \nu) \leq \\ & 2\sum_{\{\nu:\tilde{S}(\nu,n)=0\}\cap\mathcal{L}_n} \text{Prob}_{\mu_1}(\mathbf{X} \in S_0^{\nu,n}) \leq 2\text{Prob}_{\mu_1}(\mathbf{X} \in S_0^n) = 2\beta_n \end{aligned} \tag{2.8.195}$$

which certainly implies that

$$\limsup_{n\to\infty} \frac{1}{n}\log\tilde{\beta}_n \leq \limsup_{n\to\infty} \frac{1}{n}\log\beta_n \tag{2.8.196}$$

A similar computation shows that $\tilde{\alpha}_n \leq 2\alpha_n$, thus completing the proof. □

Considering from here on tests which depend only on the empirical type $L_n^{\mathbf{X}}$, the following is a characterization of an optimal rule.

Theorem 2.8.1 (Hoeffding) *Let the test S^* consist of the maps*

$$S^{*n}(\mathbf{x}) = \begin{cases} 0 & \text{if } H(L_n^{\mathbf{x}}|\mu_0) < \eta \\ 1 & \text{otherwise} \end{cases} \tag{2.8.197}$$

Then S^* is an optimal test.

Proof: By the upper bound of Theorem 2.1.1

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n}\log\text{Prob}_{\mu_0}(H_0 \text{ rejected by } S^*) &= \limsup_{n\to\infty} \frac{1}{n}\log\text{Prob}_{\mu_0}(L_n^{\mathbf{X}} \in \{\nu : H(\nu|\mu_0) \geq \eta\}) \\ &\leq -\inf_{\{\nu:H(\nu|\mu_0)\geq\eta\}} H(\nu|\mu_0) \leq -\eta \end{aligned} \tag{2.8.198}$$

Therefore, $\mathcal{S}^*$ obviously satisfies the constraint on α_n. Let β_n^* denote the β_n error probabilities associated with the test $\mathcal{S}^*$. Then, by the same upper bound (see (2.1.12) in Section 2.1)

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^* = \limsup_{n\to\infty} \frac{1}{n} \log \text{Prob}_{\mu_1}(L_n^{\mathbf{X}} \in \{\nu : H(\nu|\mu_0) < \eta\}) \leq - \inf_{\{\nu: H(\nu|\mu_0)<\eta\}} H(\nu|\mu_1) \stackrel{\Delta}{=} -I_\eta \tag{2.8.199}$$

where we have used the fact that for (2.1.12) to hold true, the set over which the minimization is performed needs not be closed. When $I_\eta = \infty$ then $\limsup_{n\to\infty} \frac{1}{n} \log \beta_n^* = -\infty$ is the best possible exponential rate. Thus, it suffices to check the optimality of $\mathcal{S}^*$ under laws μ_1 for which $I_\eta < \infty$. Fix one such law μ_1^*. Clearly $\Sigma_{\mu_1^*} \cap \Sigma_{\mu_0}$ is non-empty and moreover

$$\eta^* \stackrel{\Delta}{=} \inf_{\nu \in M_1(\Sigma_{\mu_0} \cap \Sigma_{\mu_1^*})} H(\nu|\mu_0) < \eta \,, \tag{2.8.200}$$

as $\max\{H(\nu|\mu_0), H(\nu|\mu_1^*)\} < \infty$ only when $\Sigma_\nu \subseteq \Sigma_{\mu_0} \cap \Sigma_{\mu_1^*}$. Furthermore, for any $\eta' > \eta^*$

$$I_{\eta'} \stackrel{\Delta}{=} \inf_{\{\nu: H(\nu|\mu_0)<\eta'\}} H(\nu|\mu_1^*) \;=\; \inf_{\{\nu: H(\nu|\mu_0)<\eta'\} \cap M_1(\Sigma_{\mu_0} \cap \Sigma_{\mu_1^*})} H(\nu|\mu_1^*) \;< \infty \,. \tag{2.8.201}$$

Let now $\mathcal{S}$ be any test determined by the binary function $\mathcal{S}(L_n^{\mathbf{x}}, n)$ on $M_1(\Sigma) \times Z^+$ whose error probabilities α_n satisfy the constraint (2.8.191). Then, for any $\delta > 0$ and for all $n > n_0(\delta)$ large enough

$$\mathcal{L}_n \cap \{\nu : H(\nu|\mu_0) \leq \eta - \delta\} \subset \mathcal{L}_n \cap \{\nu : \mathcal{S}(\nu, n) = 0\} \,. \tag{2.8.202}$$

For otherwise, there exists some $\delta > 0$ and a sequence of laws $\nu_n \in \mathcal{L}_n$ (for infinitely many values of n), such that $H(\nu_n|\mu_0) \leq (\eta - \delta)$ while $\mathcal{S}(\nu_n, n) = 1$. Then, by Lemma 2.1.4, for these values of n,

$$\alpha_n \geq \text{Prob}_{\mu_0}(L_n^{\mathbf{X}} = \nu_n) \geq (n+1)^{-|\Sigma|}\, e^{-nH(\nu_n|\mu_0)} \geq (n+1)^{-|\Sigma|} e^{-n(\eta-\delta)} \,,$$

implying that

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n \geq -(\eta - \delta)$$

and contradicting the constraint (2.8.191). Therefore, by combining (2.8.202) and the lower bound of (2.1.16) one obtains for any $\delta > 0$

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \beta_n \;&\geq\; \liminf_{n\to\infty} \frac{1}{n} \log \text{Prob}_{\mu_1^*}(L_n^{\mathbf{X}} \in \{\nu : H(\nu|\mu_0) < \eta - \delta\}) \\ &\geq\; -\limsup_{n\to\infty} \{ \inf_{\{\nu \in \mathcal{L}_n : H(\nu|\mu_0)<\eta-\delta\}} H(\nu|\mu_1^*)\} \stackrel{\Delta}{=} -\tilde{I}_{\eta-\delta} \end{aligned} \tag{2.8.203}$$

Since for any $k < \infty$ the set $\cup_{n=k}^{\infty} \mathcal{L}_n \cap M_1(\Sigma_{\mu_0} \cap \Sigma_{\mu_1^*})$ is a dense subset of $M_1(\Sigma_{\mu_0} \cap \Sigma_{\mu_1^*})$ over which both $H(\cdot|\mu_0)$ and $H(\cdot|\mu_1^*)$ are continuous functions. it follows by (2.8.201) that $\tilde{I}_{\eta-\delta} = I_{\eta-\delta}$ as long as $\eta - \delta > \eta^*$. Thus, one may deduce from (2.8.201) and (2.8.203) that

$$\limsup_{n\to\infty} \frac{1}{n} \log \beta_n \geq \liminf_{n\to\infty} \frac{1}{n} \log \beta_n \geq -\lim_{\delta\to 0} \tilde{I}_{\eta-\delta} = -\lim_{\delta\to 0} I_{\eta-\delta} \; . \tag{2.8.204}$$

Finally, the optimality of the test $\mathcal{S}^*$ for the law μ_1^* results by comparing (2.8.199) and (2.8.204) provided that

$$\lim_{\delta\to 0} I_{\eta-\delta} \leq I_\eta \; . \tag{2.8.205}$$

In order to prove (2.8.205) define the measure

$$\mu_0^*(a_i) \triangleq \begin{cases} \mu_0(a_i)/\mu_0(\Sigma_{\mu_1^*}) & \mu_1^*(a_i) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8.206}$$

Note that $H(\mu_0^*|\mu_0) = -\log \mu_0(\Sigma_{\mu_1^*}) = \eta^*$ (see exercise 2.8.1) and $\Sigma_{\mu_0^*} = \Sigma_{\mu_0} \cap \Sigma_{\mu_1^*}$. As $\{\nu : H(\nu|\mu_0) \leq \eta\}$ is a non-empty compact set there exists a measure ν^* such that

$$H(\nu^*|\mu_1^*) \leq I_\eta \; , \qquad H(\nu^*|\mu_0) \leq \eta$$

Consider now the family of measures $\nu_\theta = \theta\mu_0^* + (1-\theta)\nu^*$ for $\theta \in (0,1)$. Note that $\nu_\theta \in M_1(\Sigma_{\mu_0} \cap \Sigma_{\mu_1^*})$ and ν_θ converges to ν^* pointwise as $\theta \to 0$ thus implying

$$\lim_{\theta\to 0} H(\nu_\theta|\mu_1^*) = H(\nu^*|\mu_1^*) \leq I_\eta \; . \tag{2.8.207}$$

Moreover, as $H(\cdot|\mu_0)$ is a convex function

$$H(\nu_\theta|\mu_0) \leq \theta H(\mu_0^*|\mu_0) + (1-\theta)H(\nu^*|\mu_0) \leq \theta\eta^* + (1-\theta)\eta$$

implying that for any $\delta < \theta(\eta - \eta^*)$,

$$I_{\eta-\delta} \leq H(\nu_\theta|\mu_1^*) \tag{2.8.208}$$

Thus, (2.8.205) follows by combining (2.8.207) and (2.8.208). □

Remarks:

(a). The finiteness of the alphabet is essential here as (2.8.204) is obtained by applying the lower bounds of Lemma 2.1.4 for individual types instead of the natural large deviations lower

bound for *open sets* of types. Indeed, for non-finite alphabets a considerable weakening of the optimality criterion is neccessary as there are no non-trivial lower bounds for individual types (see Section ??).

(b). Both Lemma 2.8.1 and Theorem 2.8.1 may be extended to the hypothesis test problem for a known **joint law** μ_0^n versus a family of unknown **joint laws** μ_1^n provided that the random variables $X_1, \ldots, X_n$ are **finitely exchangeable** under μ_0^n and any possible μ_1^n so that the empirical measure is still a sufficient statistics. This is outlined in exercises 2.8.5-2.8.6.

Exercises:

2.8.1 Prove that for any $\nu \in M_1(\Sigma_{\mu_1^*})$, if $\nu \neq \mu_0^*$ then $H(\nu|\mu_0) > H(\mu_0^*|\mu_0) = \eta^*$.

2.8.2 Provide an alternative derivation of (2.8.202) based on the results of Section 2.7.
Hint: For any finite alphabet and any known law μ_1, deduce from (2.7.171) and exercise 2.7.1 that all probability measures within $\mathcal{L}_n$ which are of the form $\nu_\theta(a_i) \stackrel{\Delta}{=} c\mu_1(a_i)^\theta \mu_0(a_i)^{(1-\theta)}$, where $\theta \in [0,1]$ is such that $H(\nu_\theta|\mu_0) < \eta$ should eventually satisfy $\mathcal{S}(\nu_\theta, n) = 0$. Then, apply the union of probabilities bound and the volume estimate of Lemma 2.1.1.

2.8.3 (a). Let X_j for $j \notin \Delta$ be i.i.d. random variables over the finite alphabet $\Sigma = \Sigma_{\mu_0}$ while X_j for $j \in \Delta$ are *unknown deterministic points of* Σ. Prove that the test $\mathcal{S}^*$ of (2.8.197) satisfies (2.8.191) for any deterministic increasing sequence of positive integers Δ for which $\lim_{r\to\infty} \frac{\Delta_r}{r} = \infty$.
Hint: Let $L_n^{\mathbf{X}^*}$ corresponds to $\Delta = \emptyset$ and prove that $\limsup_{n\to\infty} d_V(L_n^{\mathbf{X}}, L_n^{\mathbf{X}^*}) = 0$ almost surely, with some deterministic rate of convergence which depends only upon the sequence Δ. Conclude the proof by the continuity of $H(\cdot|\mu_0)$ over $M_1(\Sigma)$.
(b). Construct a counter example to the claim above when $\Sigma_{\mu_0} \neq \Sigma$.

2.8.4 Prove that under the assumptions of exercise 2.8.3 part (a), if in addition $\Sigma_{\mu_1} = \Sigma$ for any possible μ_1 law then the test $\mathcal{S}^*$ of (2.8.197) is an optimal test. How far can you relax the assumption that $\Sigma_{\mu_1} = \Sigma$ for all μ_1 ?

2.8.5 Suppose that for any $n \in Z^+$ the random variables $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ over the finite alphabet Σ have a known joint law μ_0^n under H_0 and an unknown joint law μ_1^n under the alternative H_1. Suppose that for any n the variables $\mathbf{X}^{(n)}$ are exchangeable under both μ_0^n and any possible μ_1^n (namely, the probability of any outcome $\mathbf{X}^{(n)} = \mathbf{x}$ is invariant under permutations of indices in the vector $\mathbf{x}$). Let $L_n^{\mathbf{X}}$ denote the empirical measure (type) of $\mathbf{X}^{(n)}$ and prove that Lemma 2.8.1 holds true in this case. (Note that the variables in $\mathbf{X}^{(n)}$ may well be dependent).

2.8.6 (a). Consider the scenario described in exercise 2.8.5. Suppose that $I^n(\nu|\mu_0) = \infty$ implies $\mu_0^n(L_n^{\mathbf{X}} = \nu) = 0$ and

$$\lim_{n\to\infty} \sup_{\nu \in \mathcal{L}_n,\, I^n(\nu|\mu_0)<\infty} |\frac{1}{n} \log \mu_0^n(L_n^{\mathbf{X}} = \nu) + I^n(\nu|\mu_0)| = 0 \,. \tag{2.8.209}$$

Prove that the test $\mathcal{S}^*$ of (2.8.197) with $H(L_n^{\mathbf{X}}|\mu_0)$ replaced by $I^n(L_n^{\mathbf{X}}|\mu_0)$ is *weakly* optimal in the sense that $-\limsup_{n\to\infty}\{\frac{1}{n}\log\beta_n^*\}$ is maximal (uniformly over all possible laws μ_1^n) among all the tests for which $\limsup_{n\to\infty} \frac{1}{n}\log\alpha_n < \limsup_{n\to\infty} \frac{1}{n}\log\alpha_n^*$

(b). Apply part (a) to prove the weak optimality of thresholding $I_{\frac{n}{M},\mu_M}(L_n^{\mathbf{X}})$ of (2.1.25) when testing a given deterministic composition sequence μ_M in a sampling without replacement scheme against any unknown composition sequence for such a scheme.

2.9 Rate distortion theory for stationary and ergodic sources

Throughout this section we are interested in analyzing the following situation:

be a stationary and ergodic source, with alphabet Σ, i.e. $\mathcal{P}$ is a stationary ergodic probability measure on $\Omega = \Sigma^{Z^+}$, the space of semi-infinite sequences over Σ. Let $x_1, x_2, \ldots, x_n, \ldots$ denote an element of Ω, which we say was emitted by the source $(\mathcal{X}, \mathcal{P})$. Note that, since $\mathcal{P}$ is only ergodic, the random variables $(X_1, \cdots, X_n, \cdots)$ may well be dependent.

Next, let $\rho(x, y) : \Sigma \times \Sigma \to [0, \rho_{\max}]$ be **a one symbol bounded distortion function,** i.e. $\rho(x, x) = 0$, $\rho(x, y) \neq 0$ for $x \neq y$ and $\rho_{\max} < \infty$. The basic problem of source coding is to find a sequence of deterministic maps (codes) $C_n : \Sigma^n \to \Sigma^n$ of small distortion, where the **distortion** of a sequence of codes $\{C_n\}_{n=1}^{\infty}$ is $\limsup_{n\to\infty} \rho_{C_n}$ and

$$\rho_{C_n} \triangleq \frac{1}{n} \sum_{i=1}^{n} E[\rho(X_i, C_n(\mathbf{X})_i)] \,, \tag{2.9.210}$$

is the average distortion per symbol when the code C_n is used. The goal of the coding is to allow one to transfer the information contained in sequences emitted by the source $\mathcal{X}$ with small distortion per symbol, while transmitting as little information as possible.

Clearly, by taking C_n to be the identity map, one obtains zero distortion but with no coding gain. To get to a more meaningful situation, one would like to have a reduction in the number of possible sequences when using C_n. Let C_n also denotes the range of the map C_n and $|C_n|$ denotes the cardinality of this set. The **rate** of the code C_n is defined as

$$R_{C_n} = \frac{1}{n} \log |C_n| \tag{2.9.211}$$

and the smaller R_{C_n} is, the larger is the coding gain when using C_n. The main coding theorem, due originally to Shannon, asserts that one cannot hope to get R_{C_n} to be too small – that indeed, under a bound on the distortion, R_{C_n} is bounded below in general by some positive quantity and that there are codes which are arbitrarily close to this bound. The proof of this statement which is presented here relays on the large deviations principle of Theorem 2.3.1.

The following definitions are required for the precise statement of the coding theorem.

The distortion associated with any probability measure Q on $\Sigma \times \Sigma$ is

$$\rho_Q \stackrel{\Delta}{=} \int_{\Sigma \times \Sigma} \rho(x,y)\, dQ(x,y)\,. \tag{2.9.212}$$

Let Q_X and Q_Y be the marginals of Q. Then the **mutual information** associated with Q is

$$H(Q|Q_X \times Q_Y) \stackrel{\Delta}{=} \int_{\Sigma \times \Sigma} \log \left(\frac{dQ(x,y)}{dQ_X(x)\, dQ_Y(y)} \right) dQ(x,y) \tag{2.9.213}$$

when the above integral is well defined and finite and $H(Q|Q_X \times Q_Y) = \infty$ otherwise.[1]

The one symbol rate distortion function is defined as

$$R_1(D) = \inf_{\{Q:\rho_Q \le D, Q_X = \mathcal{P}_1\}} H(Q|Q_X \times Q_Y) \tag{2.9.214}$$

where $\mathcal{P}_1$ is the marginal on Σ of the stationary measure $\mathcal{P}$ on Σ^{Z^+}.

The one symbol distortion function $\rho(x,y)$ implies the corresponding J-symbol average distortion for $J = 2, 3, \ldots$

$$\rho^{(J)}((x_1, \ldots, x_J), (y_1, \ldots, y_J)) \stackrel{\Delta}{=} \frac{1}{J} \sum_{l=1}^{J} \rho(x_l, y_l)\,. \tag{2.9.215}$$

[1] In information theory books the mutual information is usually denoted by $I(X;Y)$. The notation $H(Q|Q_X \times Q_Y)$ is more consistent with all other notations of this book.

Thus, the J-symbol distortion associated with a probability measure Q on $\Sigma^J \times \Sigma^J$ is

$$\rho_Q^{(J)} = \int_{\Sigma^J \times \Sigma^J} \rho^{(J)}(\mathbf{x},\mathbf{y})\, dQ(\mathbf{x},\mathbf{y})\,, \tag{2.9.216}$$

and the *mutual information* associated with the measure Q having marginals Q_X, Q_Y on Σ^J is

$$H(Q|Q_X \times Q_Y) \stackrel{\Delta}{=} \int_{\Sigma^J \times \Sigma^J} \log\left(\frac{dQ(\mathbf{x},\mathbf{y})}{dQ_X(\mathbf{x})\, dQ_Y(\mathbf{y})}\right) dQ(\mathbf{x},\mathbf{y})\,. \tag{2.9.217}$$

The J-symbol rate distortion function is therefore defined as

$$R_J(D) \stackrel{\Delta}{=} \inf_{\{Q:\rho_Q^{(J)} \le D,\, Q_X = \mathcal{P}_J\}} \frac{1}{J} H(Q|Q_X \times Q_Y)\,, \tag{2.9.218}$$

where $\mathcal{P}_J$ is the J–th marginal (on Σ^J) of the stationary measure $\mathcal{P}$. Finally, the **rate distortion function** is

$$R(D) \stackrel{\Delta}{=} \inf_{J \ge 1} R_J(D)\,. \tag{2.9.219}$$

The source coding theorem states that the rate distortion function is a tight lower bound on the limiting rate R_{C_n} of a sequence of codes $\{C_n\}_{n=1}^{\infty}$ with distortion D.

Theorem 2.9.1 (Source Coding Theorem)

(a). **Direct Part:** *For any $D \ge 0$ such that $R(D) < \infty$ and any $\delta > 0$, there exists a sequence of codes $\{C_n\}_{n=1}^{\infty}$ with distortion at most D and rates $R_{C_n} \le R(D) + \delta$.*

(b). **Converse Part:** *For any sequence of codes $\{C_n\}_{n=1}^{\infty}$ of distortion D and any $\delta > 0$ $\liminf_{n\to\infty} R_{C_n} \ge R(D+\delta)$.*

Remark: Note that $|\Sigma|$ may be infinite and there are no structural conditions on Σ besides the requirement that $\mathcal{P}$ be based on Σ^{Z^+}. On the other hand, whenever $R(D)$ is finite, the resulting codes always take values in some finite set and in particular, may be represented by finite binary sequences.

The proof of the Source Coding Theorem is presented via a sequence of lemmas with the large deviations principle of Section 2.3 implying the first lemma which is key for the Direct Part of the theorem.

Lemma 2.9.1 *Suppose Q is any probability measure on $\Sigma \times \Sigma$ for which $I_Q \stackrel{\Delta}{=} H(Q|Q_X \times Q_Y) < \infty$, $Q_X \equiv \mathcal{P}_1$ and $\rho_Q < \rho_{Q_X \times Q_Y}$ (where $\rho_{Q_X \times Q_Y}$ is the distortion associated with $Q_X \times Q_Y \equiv \mathcal{P}_1 \times$*

Q_Y). Let $Z_n(\mathbf{x}) \stackrel{\Delta}{=} \frac{1}{n}\sum_{j=1}^{n} \rho(x_j, Y_j)$ where Y_j are independent random variables distributed over Σ according to Q_Y. Then,

$$\liminf_{n\to\infty} \frac{1}{n} \log \text{Prob}(Z_n(\mathbf{X}) < \rho_Q | \mathbf{X}) \geq -I_Q \quad \textit{almost surely } \mathcal{P} \tag{2.9.220}$$

where $\mathbf{X}$ is a random sequence of symbols emitted by the source $\mathcal{X}$.

Proof: The inequality (2.9.220) follows from the lower bound of Theorem 2.3.1 for the open set $\Gamma = (-\infty, \rho_Q)$ and the random variables $\{Z_n\}_{n=1}^{\infty}$. Specifically, Theorem 2.3.1 is applied per element of $\Omega = \Sigma^{Z^+}$, where the conditions of this theorem hold almost surely by the ergodicity of $\mathcal{P}$. To verify that, let $\Lambda_n(\theta) \stackrel{\Delta}{=} \log E[e^{\theta Z_n(\mathbf{X})}|\mathbf{X}]$ and note that since $\rho(\cdot,\cdot)$ is a bounded function $|\Lambda_n(\theta)| < \infty$ for all $\theta \in \mathbb{R}$ and any realization of $\mathbf{X}$. By Birckhoff's ergodic theorem [3]

$$\Lambda(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} \Lambda_n(n\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \log \int_{\Sigma} e^{\theta\rho(X_j, y)} dQ_Y(y) \tag{2.9.221}$$

exists almost surely $\mathcal{P}$. Moreover,

$$\Lambda(\theta) = \int_{\Sigma} \log \left(\int_{\Sigma} e^{\theta\rho(x,y)} dQ_Y(y) \right) d\mathcal{P}_1(x) = \int_{\Sigma} \log \left(\int_{\Sigma} e^{\theta\rho(x,y)} dQ_Y(y) \right) dQ_X(x)\,, \tag{2.9.222}$$

does not depend on the specific sequence $\mathbf{X}$ emitted by the source (recall that $Q_X \equiv \mathcal{P}_1$ implying the second equality above). Furthermore, since $\rho(\cdot,\cdot)$ is uniformly bounded the function $\Lambda(\cdot)$ is finite and differentiable everywhere (in $\mathbb{R}^1$). Therefore, by Lemma 2.3.2, part (c) of Theorem 2.3.1 applies, yielding

$$\liminf_{n\to\infty} \frac{1}{n} \log P(Z_n(\mathbf{X}) < \rho_Q | \mathbf{X}) \geq -\inf_{x<\rho_Q} \Lambda^*(x) \stackrel{\Delta}{=} -J_Q \quad \text{almost surely} \quad \mathcal{P} \tag{2.9.223}$$

Recall that $\Lambda^*(x) \stackrel{\Delta}{=} \sup_{\lambda}(\lambda x - \Lambda(\lambda))$. As $\Lambda(0) = 0$ (see (2.9.222)), $\Lambda'(0) = \rho_{Q_X \times Q_Y}$ (differentiate (2.9.222) and compare with (2.9.212)) and $\Lambda'(\lambda)$ is monotonically nondecreasing (since $\rho(\cdot,\cdot) \geq 0$) it follows that for $x \leq \rho_Q < \Lambda'(0)$

$$\sup_{\lambda>0}[\lambda x - \Lambda(\lambda)] \leq \sup_{\lambda>0}[\lambda\Lambda'(0) - \Lambda(\lambda)] \leq \Lambda^*(\Lambda'(0)) = 0\Lambda'(0) - \Lambda(0) = 0\,. \tag{2.9.224}$$

Since $\Lambda^*(x) \geq 0$, it follows that for $x \leq \rho_Q$

$$\Lambda^*(x) = \sup_{\lambda\leq 0}[\lambda x - \Lambda(\lambda)] \geq \Lambda^*(\rho_Q) \tag{2.9.225}$$

and thus $J_Q = \Lambda^*(\rho_Q)$.

For any $\lambda \in \mathbb{R}^1$ define the probability measure Q_λ on $\Sigma \times \Sigma$ via

$$\frac{dQ_\lambda(x,y)}{dQ_X(x)dQ_Y(y)} \triangleq \frac{e^{\lambda\rho(x,y)}}{\int_\Sigma e^{\lambda\rho(x,z)}dQ_Y(z)}.$$

Since $\rho(\cdot,\cdot)$ is a bounded function the measures Q_λ and $Q_X \times Q_Y$ are equivalent and as $H(Q|Q_X \times Q_Y) < \infty$ the relative entropy $H(Q|Q_\lambda)$ is well defined. Moreover,

$$\begin{aligned} 0 \quad &\leq H(Q|Q_\lambda) \triangleq \int_{\Sigma\times\Sigma} \log \frac{dQ(x,y)}{dQ_\lambda(x,y)} dQ(x,y) = \\ &= \int_{\Sigma\times\Sigma} \log \left\{ \frac{dQ(x,y)e^{-\lambda\rho(x,y)}}{dQ_X(x)dQ_Y(y)} \int_\Sigma e^{\lambda\rho(x,z)}dQ_Y(z) \right\} dQ(x,y) = I_Q - \lambda\rho_Q + \Lambda(\lambda) \end{aligned} \quad (2.9.226)$$

Since (2.9.226) holds for all λ, it follows that $I_Q \geq \Lambda^*(\rho_Q)$. The proof is now completed in view of (2.9.223) and (2.9.225). □

The proof of the Direct Part of Theorem 2.9.1 is based on a random coding argument, where instead of specifically constructing the codes C_n, the classes $\mathcal{C}_n$ of all codes of some fixed size are considered. Let $\overline{\rho}_n \triangleq E_{\mathcal{C}_n}[\rho_{C_n}]$ be the average of ρ_{C_n} over $\mathcal{C}_n$, where the distribution within the class $\mathcal{C}_n$ results by choosing the code C_n at random according to some probability measure. For any probability measure over $\mathcal{C}_n$ there exists at least one code in $\mathcal{C}_n$ for which $\rho_{C_n} \leq \overline{\rho}_n$. In the following lemma an upper bound on $\overline{\rho}_n$ is derived based on the large deviations lower bound of Lemma 2.9.1.

Lemma 2.9.2 *Suppose Q is a probability measure on $\Sigma \times \Sigma$ for which $H(Q|Q_X \times Q_Y) < \infty$ and $Q_X \equiv \mathcal{P}_1$. Fix $\delta > 0$ arbitrarily small and let $\mathcal{C}_n$ be the class of all codes C_n of size $|C_n| = \lfloor e^{n(H(Q|Q_X\times Q_Y)+\delta)} \rfloor$. Then, there exist distributions on $\mathcal{C}_n$ for which $\limsup_{n\to\infty} \overline{\rho}_n \leq \rho_Q$.*

Proof: Fix the probability measure Q and let $I_Q \triangleq H(Q|Q_X \times Q_Y)$.

(a). Suppose that $\rho_{Q_X\times Q_Y} \leq \rho_Q$ (recall that $\rho_{Q_X\times Q_Y}$ is the distortion associated with the measure $Q_X \times Q_Y \equiv \mathcal{P}_1 \times Q_Y$). Let $Y_1, \ldots, Y_n$ be i.i.d according to the law Q_Y and $\mathbf{Y} \triangleq (Y_1, \ldots, Y_n)$ be a code-word of C_n to which all Σ^n is mapped. Since $I_Q \geq 0$ this construction is always possible ($|C_n| \geq 1$) and it results with a class of codes $\mathcal{C}_n$, one code per realization of $\mathbf{Y}$. The average distortion over codes in $\mathcal{C}_n$ is now

$$\overline{\rho}_n = E_{\mathbf{Y}} \left[\frac{1}{n} \sum_{j=1}^{n} E[\rho(X_j, Y_j)|\mathbf{Y}] \right] = \int_{\Sigma\times\Sigma} \rho(x,y)\, d\mathcal{P}_1(x) dQ_Y(y) = \rho_{Q_X\times Q_Y}. \quad (2.9.227)$$

Thus, since $\overline{\rho}_n = \rho_{Q_X \times Q_Y} \leq \rho_Q$ the proof of the lemma is complete.

(b). Consider now the case when $\rho_{Q_X \times Q_Y} > \rho_Q$, $I_Q < \infty$ and $Q_X \equiv \mathcal{P}_1$. Let $k_n \stackrel{\Delta}{=} |C_n| = \lfloor e^{n(I_Q+\delta)} \rfloor$ be as specified in the statement of the lemma and $Y_j^{(i)}$, $j = 1, \ldots, n$, $i = 1, \ldots, k_n$, be $n \times k_n$ i.i.d. random variables of law Q_Y. The probability distribution on the class of codes $\mathcal{C}_n$ is generated by considering the codes with code-words $\mathbf{Y}^{(i)} \stackrel{\Delta}{=} (Y_1^{(i)}, \ldots, Y_n^{(i)})$ for $i = 1, \ldots, k_n$. Per realization of these code-words the mapping C_n is constructed as follows. For any $\mathbf{x} \in \Sigma^n$ define the set

$$S_n(\mathbf{x}) \stackrel{\Delta}{=} \{\mathbf{y} : \frac{1}{n} \sum_{j=1}^{n} \rho(x_j, y_j) < \rho_Q\} \tag{2.9.228}$$

and let $C_n(\mathbf{x})$ be any element of $C_n \cap S_n(\mathbf{x})$, where if this set is empty then $C_n(\mathbf{x})$ is arbitrarily chosen. For this mapping,

$$\frac{1}{n} \sum_{j=1}^{n} \rho(x_j, C_n(\mathbf{x})_j) \leq \rho_Q + \rho_{\max} 1_{C_n \cap S_n(\mathbf{x}) = \emptyset}$$

implying that

$$\overline{\rho}_n \leq \rho_Q + \rho_{\max} \mathrm{Prob}(C_n \cap S_n(\mathbf{X}) = \emptyset) \,, \tag{2.9.229}$$

where $\mathbf{X} \in \Sigma^n$ is a random sequence of n symbols emitted by the source $\mathcal{X}$ and the set C_n consists of the k_n i.i.d. random vectors $\mathbf{Y}^{(i)}$. Clearly,

$$\begin{aligned} &\mathrm{Prob}(C_n \cap S_n(\mathbf{X}) = \emptyset) = E[\mathrm{Prob}(\mathbf{Y}^{(i)} \notin S_n(\mathbf{X}) \text{ for all } i | \mathbf{X})] \\ = \;& E[(1 - \mathrm{Prob}(\mathbf{Y}^{(1)} \in S_n(\mathbf{X})|\mathbf{X}))^{k_n}] \leq E[e^{-k_n \mathrm{Prob}(\mathbf{Y}^{(1)} \in S_n(\mathbf{X})|\mathbf{X})}] \,. \end{aligned} \tag{2.9.230}$$

By (2.9.229) and (2.9.230), $\limsup_{n \to \infty} \overline{\rho}_n \leq \rho_Q$ provided that $k_n \mathrm{Prob}(\mathbf{Y}^{(1)} \in S_n(\mathbf{X})|\mathbf{X}) \to \infty$ as $n \to \infty$, in probability $\mathcal{P}$. By the definition of k_n it suffices to show that

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathrm{Prob}(\mathbf{Y}^{(1)} \in S_n(\mathbf{X})|\mathbf{X}) \geq -I_Q \quad \text{in Probability } \mathcal{P} \,, \tag{2.9.231}$$

in order to complete the proof of the lemma. Since Lemma 2.9.1 applies to $Z_n(\mathbf{x}) \stackrel{\Delta}{=} \frac{1}{n} \sum_{j=1}^{n} \rho(x_j, Y_j^{(1)})$ and $\{\mathbf{Y}^{(1)} \in S_n(\mathbf{x})\} \equiv \{Z_n(\mathbf{x}) < \rho_Q\}$ the bound (2.9.231) follows. □

The following weak version of the direct part of the Source Coding Theorem is an immediate consequence of Lemma 2.9.2.

Lemma 2.9.3 *For any $D \geq 0$ such that $R_1(D) < \infty$ and any $\delta > 0$ there exists a sequence of codes $\{C_n\}_{n=1}^{\infty}$ with distortion at most D and rates $R_{C_n} \leq R_1(D) + \delta$.*

Proof: Since $R_1(D) < \infty$ there exists a sequence of measures $\{Q^{(m)}\}_{m=1}^{\infty}$ such that $I_{Q^{(m)}} \to R_1(D)$ while $\rho_{Q^{(m)}} \leq D$ and $Q_X^{(m)} \equiv \mathcal{P}_1$. By applying Lemma 2.9.2 for $Q^{(m)}$, $m = 1, 2, \cdots$, it follows that $\limsup_{n\to\infty} \overline{\rho}_n \leq \limsup_{m\to\infty} \rho_{Q^{(m)}} \leq D$ when $\mathcal{C}_n$ is the class of all codes of size $\lfloor e^{n(R_1(D)+\delta)} \rfloor$ and $\delta > 0$ is arbitrarily small. The existence of a sequence of codes C_n of rates $R_{C_n} \leq R_1(D) + \delta$ and of distortion $\limsup_{n\to\infty} \rho_{C_n} \leq D$ is deduced by extracting the codes C_n of minimal ρ_{C_n} from the ensembles $\mathcal{C}_n$. □

For i.i.d. source symbols, $R(D) = R_1(D)$ (see exercise 2.9.2) and the Direct Part of the Source Coding Theorem amounts to Lemma 2.9.3. When the symbols are dependent one needs the following extension of Lemma 2.9.2.

Lemma 2.9.4 *Suppose $\mathcal{P}$ is ergodic with respect to the J-th shift operation (namely, it is ergodic in blocks of size J). Then, for any probability measure Q on $\Sigma^J \times \Sigma^J$ with $Q_X \equiv \mathcal{P}_J$ and for any $\delta > 0$ there exists a sequence of codes C_n of rates $R_{C_n} \leq \frac{1}{J} H(Q|Q_X \times Q_Y) + \delta$ and of distortion at most $\rho_Q^{(J)}$.*

Proof: Consider the enlarged alphabet Σ^J and regard each block of J consecutive symbols of the emitted source sequence as one symbol from Σ^J. A sketch of the proof which basically follows the proofs of Lemmas 2.9.1 and 2.9.2 is presented here. Let

$$\Lambda^{(J)}(\theta) = \frac{1}{J} \int_{\Sigma^J} \log \left(\int_{\Sigma^J} e^{J\theta\rho^{(J)}(x,y)} dQ_Y(y) \right) d\mathcal{P}_J(x) \,. \tag{2.9.232}$$

Then, by the ergodicity of $\mathcal{P}$ in blocks of size J

$$\Lambda^{(J)}(\theta) = \lim_{n\to\infty} \frac{1}{nJ} \log E[e^{n\theta J Z_{nJ}^{(J)}(\mathbf{X})} | \mathbf{X}] \quad \text{almost surely } \mathcal{P} \,, \tag{2.9.233}$$

where $Z_{nJ}^{(J)}(\mathbf{x}) \stackrel{\Delta}{=} \frac{1}{n} \sum_{j=1}^{n} \rho^{(J)}(\mathbf{x}_j, \mathbf{Y}_j)$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ are concatenated symbols namely, are elements of Σ^J (while $\mathbf{Y}_1, \ldots \mathbf{Y}_n$ are i.i.d. random variables of law Q_Y over Σ^J). Thus, Theorem 2.3.1 is once again applicable. To complete the proof, define Q_λ now via

$$\frac{dQ_\lambda(\mathbf{x}, \mathbf{y})}{dQ_X(\mathbf{x}) dQ_Y(\mathbf{y})} \stackrel{\Delta}{=} \frac{e^{J\lambda\rho^{(J)}(\mathbf{x},\mathbf{y})}}{\int_{\Sigma^J} e^{J\lambda\rho^{(J)}(\mathbf{x},\mathbf{z})} dQ_Y(\mathbf{z})} \,,$$

and observe that for any $\lambda \in \mathbb{R}^1$ and any probability measure Q over $\Sigma^J \times \Sigma^J$ with $Q_X \equiv \mathcal{P}_J$

$$0 \leq \frac{1}{J} H(Q|Q_\lambda) = \frac{1}{J} H(Q|Q_X \times Q_Y) - \lambda \rho_Q^{(J)} + \Lambda^{(J)}(\lambda) \,.$$

Thus, $\frac{1}{J}H(Q|Q_X \times Q_Y) \geq \Lambda^{(J)*}(\rho_Q^{(J)})$. □

The following corollary is an immediate consequence of Lemma 2.9.4 (by adapting the proof of Lemma 2.9.3 to $J \neq 1$).

Corollary 2.9.1 *If $\mathcal{P}$ is ergodic in blocks of size J for any $J \in Z^+$ then the direct part of the Source Coding Theorem holds.*

While in general, an ergodic $\mathcal{P}$ might be non ergodic in blocks, Corollary 2.9.1 holds true for any stationary and ergodic $\mathcal{P}$ (see for example [2], pp. 278–280, [13], pp. 496–500). It clearly suffices for that purpose to prove the following lemma.

Lemma 2.9.5 *Lemma 2.9.4 holds true for any stationary and ergodic $\mathcal{P}$ and any $J \in Z^+$.*

Proof: It is possible to show that when considering blocks of size J, the emitted infinite sequences of the source may almost surely be divided into J equally probable ergodic modes, $E_0, \ldots, E_{J-1}$, such that if sequence $(x_1, x_2, \cdots, x_n, \cdots)$ belongs to mode E_i then $(x_{1+k}, x_{2+k}, \cdots, x_{n+k}, \cdots)$ belongs to the mode $E_{(i+k) \bmod J}$ (see [2],[13]). This implies that $\mathcal{P} = \frac{1}{J}\sum_{i=0}^{J-1} \mathcal{P}^{(i)}$ where $\{\mathcal{P}^{(i)}\}_{i=0}^{J-1}$ are ergodic in blocks of size J and correspond to the ergodic modes E_i. By projections onto these ergodic modes, each law Q on $\Sigma^J \times \Sigma^J$ with $Q_X \equiv \mathcal{P}_J$ is similarly decomposed into $Q = \frac{1}{J}\sum_{i=0}^{J-1} Q^{(i)}$ such that $Q_X^{(i)} = \mathcal{P}_J^{(i)}$ is the J-th marginal of $\mathcal{P}^{(i)}$ while $\rho_Q^{(J)} = \frac{1}{J}\sum_{i=0}^{J-1} \rho_{Q^{(i)}}^{(J)}$ and

$$H(Q|Q_X \times Q_Y) \geq \frac{1}{J}\sum_{i=0}^{J-1} H(Q^{(i)}|Q_X^{(i)} \times Q_Y^{(i)}) \,. \tag{2.9.234}$$

By applying Lemma 2.9.4 for $\{Q^{(i)}\}_{i=0}^{J-1}$ there exist J sequences of codes $C_n^{(i)}$, $i = 0, \ldots, J-1$ with rates $R_{C_n^{(i)}} \leq \frac{1}{J}H(Q^{(i)}|Q_X^{(i)} \times Q_Y^{(i)}) + \delta$ and distortions at most $\rho_{Q^{(i)}}^{(J)}$ *with respect to the source measures* $\{\mathcal{P}^{(i)}\}_{i=0}^{J-1}$, respectively. A sequence of codes $C_{(n+1)J}^*$ of rates

$$R_{C_{(n+1)J}^*} \leq \frac{1}{J}\sum_{i=0}^{J-1} R_{C_n^{(i)}} + \frac{1}{(n+1)J}\log J \tag{2.9.235}$$

and of distortion at most $\rho_Q^{(J)}$ with respect to $\mathcal{P}$ is now constructed as follows. The code $C_{(n+1)J}^*$ is the union of J codes $\{\tilde{C}_{(n+1)J}^{(i)}\}_{i=0}^{J-1}$, each of cardinality $\prod_{i=0}^{J-1}|C_n^{(i)}|$ and length $(n+1)J$. The code $\tilde{C}_{(n+1)J}^{(i)}$ consists of all the distinct words in

$\{(y_0, a^*, y_1, a^*, \cdots, y_{J-1}, a^*) : y_k \in C_n{}^{(i+k) \bmod J}, k = 0, \ldots, J-1\}$ and $a^* \in \Sigma$ is a fixed separator symbol. This construction guarantees that the sequence of codes $\tilde{C}^{(i)}_{(n+1)J}$ has distortion at most $\rho_Q^{(J)}$ with respect to the source measure $\mathcal{P}^{(i)}$ and thus $C^*_{(n+1)J}$ has at most this distortion with respect to the source measure $\mathcal{P}$. By (2.9.234) and (2.9.235) for all n large enough, $R_{C^*_{(n+1)J}} \leq \frac{1}{J} H(Q|Q_X \times Q_Y) + 2\delta$. Finally, for code length which is not an integer multiple of J, one may easily modify the code $C^*_{(n+1)J}$ of closest length while neither affecting the *limiting* distortion per symbol nor the limiting rate (as $n \to \infty$). The proof is thus complete. □

The last lemma in this section is devoted to the converse part of the Source Coding Theorem, whose proof is based on information theoretical arguments and not on large deviations bounds.

Lemma 2.9.6 *For any sequence of codes $\{C_n\}_{n=1}^{\infty}$ of distortion D and for any $\delta > 0$* $\liminf_{n\to\infty} R_{C_n} \geq R(D+\delta)$.

Proof: It suffices to consider codes C_n of finite rates and of distortion D. Such a code C_n is a mapping from Σ^n to Σ^n. When its domain Σ^n is equipped with the probability measure $\mathcal{P}_n$ (the n-th marginal of the source measure $\mathcal{P}$), C_n induces a (degenerate) joint measure $Q^{(n)}$ on $\Sigma^n \times \Sigma^n$. Note that $Q_X^{(n)} \equiv \mathcal{P}_n$ and $\rho_{Q^{(n)}} = \rho_{C_n} \leq D + \delta$ for all n large enough (and any $\delta > 0$). Therefore, $H(Q^{(n)}|Q_X^{(n)} \times Q_Y^{(n)}) \geq nR_n(D+\delta) \geq nR(D+\delta)$ for any $\delta > 0$ and any n large enough. Since the marginals $Q_Y^{(n)}$ have the finite support sets C_n

$$nR_{C_n} = \log |C_n| \geq H(Q_Y^{(n)}),$$

where the entropy $H(Q_Y^{(n)})$ is defined as $H(Q_Y^{(n)}) \triangleq \sum_{i=1}^{|C_n|} Q_Y^{(n)}(y_i) \log \frac{1}{Q_Y^{(n)}(y_i)}$ (compare with the definition in Section 2.1.1). Let $f_n(x, y_i) \triangleq \frac{dQ^{(n)}(x,y_i)}{dQ_X^{(n)}(x) dQ_Y^{(n)}(y_i)}$. Then, $f_n : \Sigma^n \times \Sigma^n \to [0, \infty)$ is well defined, $\int_{\Sigma^n} f_n(x, y_i)\, dQ_X^{(n)}(x) = 1$ for all i as well as $\sum_{i=1}^{|C_n|} f_n(x, y_i)\, Q_Y^{(n)}(y_i) = 1$ almost surely $Q_X^{(n)}$. Thus, $f_n(x, y_i)\, Q_Y^{(n)}(y_i) \leq 1$ almost surely $Q_X^{(n)}$ and

$$H(Q_Y^{(n)}) - H(Q^{(n)}|Q_X^{(n)} \times Q_Y^{(n)}) = \sum_{i=1}^{|C_n|} Q_Y^{(n)}(y_i) \left[\log \frac{1}{Q_Y^{(n)}(y_i)} - \int_{\Sigma^n} f_n(x, y_i) \log f_n(x, y_i)\, dQ_X^{(n)}(x) \right]$$

$$= \int_{\Sigma^n} dQ_X^{(n)}(x) \left\{ \sum_{i=1}^{|C_n|} f_n(x, y_i) Q_Y^{(n)}(y_i) \log \left[\frac{1}{f_n(x, y_i)\, Q_Y^{(n)}(y_i)} \right] \right\} \geq 0 \qquad (2.9.236)$$

Therefore,

$$\liminf_{n\to\infty} R_{C_n} \geq \liminf_{n\to\infty} \frac{1}{n} H(Q_Y^{(n)}) \geq \liminf_{n\to\infty} \frac{1}{n} H(Q^{(n)}|Q_X^{(n)} \times Q_Y^{(n)}) \geq R(D+\delta)\,.$$

□

Exercises:

2.9.1 (a). Prove that when Σ is a finite set and $R_1(D) > 0$, there exists a probability measure Q on $\Sigma \times \Sigma$ for which $I_Q = R_1(D)$, $Q_X = \mathcal{P}_1$ and $\rho_Q = D$.
(b). Prove that for this measure also $\Lambda^*(\rho_Q) = I_Q$.

2.9.2 Prove that when $\mathcal{P}$ is a product measure (namely, the emitted symbols $X_1, X_2, \cdots, X_n$ are i.i.d.) then $R_J(D) = R_1(D)$ for all $J \in Z^+$.

2.9.3 (a). Show that $(m+n)R_{m+n}(D) \leq mR_m(D) + nR_n(D)$ for any two integers m, n.
(b). Conclude that if $\limsup_{J\to\infty} R_J(D) < \infty$ then $R(D) = \lim_{J\to\infty} R_J(D)$.

emphasize in intr. that deal with iid case. correctiuons are denoted

2.10 Refinements of large deviations statements in $\mathbb{R}^d$

Cramer's theorem deals with the tails of the empirical means $\hat{S}_n$. On a finer scale, at least in the i.i.d. case, the random variables $\sqrt{n}\hat{S}_n$ possess a limiting Normal distribution by the classical Central Limit Theorem. In this situation, the empirical means $n^\beta \hat{S}_n$ satisfy a large deviations principle for any $\beta \in (0, \frac{1}{2})$, but always with a quadratic (Normal-like) rate function. This statement is made precise in the following theorem.

Theorem 2.10.1 *Let $X_1, \ldots, X_n$ be a sequence of $\mathbb{R}^d$ valued i.i.d. random variables with $E(X_i) = 0$ such that $\Lambda_X(\lambda) \stackrel{\Delta}{=} \log E[e^{<\lambda, X_i>}] < \infty$ in some open ball $B_{0,\delta}$ around the origin. Fix $\beta \in (0, \frac{1}{2})$ and let $Z_n \stackrel{\Delta}{=} \frac{1}{n^{1-\beta}} \sum_{i=1}^n X_i = n^\beta \hat{S}_n$. Then, Z_n satisfy a large deviations principle in $\mathbb{R}^d$ governed by the good rate function*

$$I_g(x) \stackrel{\Delta}{=} \frac{1}{2} < x, \mathbf{C}x > , \tag{2.10.237}$$

where $\mathbf{C}$ is the covariance matrix of X_i. Moreover, any open or closed set G is an I_g continuity set.

Proof: This theorem follows from the general large deviations statement of Section 2.3 with $a_n \stackrel{\Delta}{=} n^{(2\beta-1)}$. Indeed, in the notations of Section 2.3

$$\begin{aligned}\Lambda_n(a_n^{-1}\lambda) &= \log E\left(e^{a_n^{-1}<\lambda,Z_n>}\right) \\ &= \sum_{i=1}^{n} \log E\left(e^{n^{-\beta}<\lambda,X_i>}\right) = n \log E\left(e^{n^{-\beta}<\lambda,X_1>}\right) \end{aligned} \tag{2.10.238}$$

Therefore

$$\Lambda(\lambda) = \lim_{n\to\infty} n^{2\beta} \log E\left(e^{n^{-\beta}<\lambda,X_1>}\right) \tag{2.10.239}$$

Note that $n^{-\beta} \underset{n\to\infty}{\longrightarrow} 0$ and therefore, by our assumption that $\Lambda(\lambda) < \infty$ in an open ball around 0, for each $\lambda \in \mathbb{R}^d$ there exists an n_0 large enough such that for all $n > n_0$, $E\left(e^{n^{-\beta}<\lambda,X_1>}\right) < \infty$, and by dominated convergence

$$E\left(e^{n^{-\beta}<\lambda,X_1>}\right) = 1 + n^{-\beta} E[<\lambda, X_1>] + \frac{1}{2} n^{-2\beta} E[<\lambda, X_1>^2] + \mathbf{O}\left(n^{-3\beta}\right) \tag{2.10.240}$$

Substituing (2.10.240) into (2.10.239) one obtains (using the identity $E[<\lambda, X_1>] = 0$)

$$\begin{aligned}\Lambda(\lambda) &= \lim_{n\to\infty} n^{2\beta} \log\left\{1 + \frac{1}{2} n^{-2\beta} E[<\lambda, X_1>^2] + \mathbf{O}(n^{-3\beta})\right\} \\ &= \frac{1}{2} E[<\lambda, X_1>^2] = \frac{1}{2} <\lambda, \mathbf{C}\lambda>^2 \end{aligned} \tag{2.10.241}$$

Thus,

$$\Lambda^*(x) \stackrel{\Delta}{=} \sup_{\lambda\in\mathbb{R}^d} \{<\lambda, x> - \Lambda(\lambda)\} = \sup_{\lambda\in\mathbb{R}^d} \{<\lambda, x> - \frac{1}{2} <\lambda, \mathbf{C}\lambda>\} = \frac{1}{2} <x, \mathbf{C}x> = I_g(x) \tag{2.10.242}$$

Since $\Lambda^*(\lambda) = \frac{1}{2} <\lambda, \mathbf{C}\lambda>$ is differentiable and finite everywhere Theorem 2.3.1 applies and the proof is thus complete. □

Remarks:

(a). Note that Theorem 2.10.1 is nothing more than what one would obtain from a naive Taylor expansion applied on the anastaz $\text{Prob}(\hat{S}_n = x) \approx e^{-nI(x)}$ where $I(\cdot)$ is the rate function of Theorem 2.3.2 (see Section 2.3).

(b). The rate of convergence in (2.10.241) is of $\mathbf{O}(n^{-\beta})$, suggesting a similar convergence rate for $a_n \log \text{Prob}(Z_n \in G)$ (which converges to $-\inf_{x\in G} I_g(x)$ for any open G). Indeed, such a result is proved in [12] (pp. 552-553) for $d = 1$ and $G = (x, \infty)$. This extends for example the validity of

the Normal approximation for the distribution of $\sqrt{n}\hat{S}_n$ to intervals of $o(n^{\frac{1}{6}})$ (which correspond to $\beta > \frac{1}{3}$ here).

(c). A similar result may be obtained in the context of Markov additive processes (see Section 2.4.1).

Another refinement of Cramer's theorem involves a more accurate estimate of the laws μ_n of $\hat{S}_n$ (the empirical means of i.i.d. random variables). Specifically, for a "nice" I Continuity Set A one seeks an estimate J_n of $\mu_n(A)$ such that $\lim_{n\to\infty} J_n\mu_n(A) = 1$. Such an estimate is an improvement over the normalized logarithmic limit $\frac{1}{n}\log\mu_n(A)$ implied by a large deviations principle. The following theorem deals with the estimate J_n for certain half intervals $A = [q,\infty) \subset \mathbb{R}^1$.

Theorem 2.10.2 (Bahadur and Rao) *Let μ_n denotes the law of $\hat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$ where X_i are i.i.d. real valued random variables and $\Lambda(\lambda) = \log E[e^{\lambda X_1}]$ is the logarithmic moment generating function of X_i. Consider the set $A = [q,\infty)$ where $q \in \mathcal{F}$, namely $q = \Lambda'(\eta)$ for some positive $\eta \in \mathcal{D}_\Lambda^o$.*

(a). If the law of X_1 is non-lattice, then

$$\lim_{n\to\infty} J_n\mu_n(A) = 1 \tag{2.10.243}$$

where $J_n = \eta\sqrt{\Lambda''(\eta)\,2\pi n}\; e^{n\Lambda^*(q)}$.

(b). Suppose X_1 has a lattice law, namely, for some finite x_0, d, the random variable $\frac{1}{d}(X_1 - x_0)$ is with probability one an integer value. Assume further that $1 > \mathrm{Prob}\,(X_1 = q) > 0$ (in particular, this implies that $\frac{1}{d}(q - x_0)$ is an integer and that $\Lambda''(\eta) > 0$). Then,

$$\lim_{n\to\infty} J_n\mu_n(A) = \frac{\eta d}{1 - e^{-\eta d}}\,. \tag{2.10.244}$$

Remarks: (a). Recall that $\Lambda^*(q) = \eta q - \Lambda(\eta)$ and $\Lambda(\cdot)$ is C^∞ in some open neighborhood of η by the dominated convergence argument (for details see Lemma 2.2.1 and exercise 2.2.5).

(b). Actually the limit relations (2.10.243) and (2.10.244) remain valid even for small intervals of size of $O(\frac{\log n}{n})$ (see exercise 2.10.1).

(c). The proof of this theorem is based on an exponential translation of a local Central Limit Theorem. This approach is applicable for the dependent case of Section 2.3 and to certain extent applies also in $\mathbb{R}^d$, $d > 1$.

Proof: (a). Consider the probability measure $\tilde{\mu}$ defined by $d\tilde{\mu}(x) \triangleq e^{\eta x - \Lambda(\eta)} d\mu(x)$ and let $Y_i \triangleq (X_i - q)/\sqrt{\Lambda''(\eta)}$, for $i = 1, 2, \ldots, n$. Note that $Y_1, \ldots, Y_n$ are i.i.d. random variables with $E_{\tilde{\mu}}[Y_1] = 0$ and $E_{\tilde{\mu}}[Y_1^2] = 1$ (this can be easily checked by computing the first two moments of X_1 under $\tilde{\mu}$). Let $F_n(x)$ denotes the distribution function of $W_n \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ under the measure $\tilde{\mu}$.

Since X_i are non-lattice, the Berry-Esseen expansion of $F_n(x)$ results with:

$$\lim_{n\to\infty} \left\{ \sqrt{n} \sup_x \left| F_n(x) - \Phi(x) - \frac{m_3}{6\sqrt{n}} (1 - x^2)\, \phi(x) \right| \right\} = 0\,, \tag{2.10.245}$$

where $m_3 \triangleq E_{\tilde{\mu}}[Y_1^3] < \infty$, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is the standard Normal density, and $\Phi(x) = \int_{-\infty}^x \phi(\eta) d\eta$ (for the derivation of (2.10.245) see [12] page 512).

Now,

$$\mu_n(A) = \mu_n([q, \infty)) = E_{\tilde{\mu}}[e^{-n[\eta \hat{S}_n - \Lambda(\eta)]} 1_{\hat{S}_n \geq q}] =$$

$$e^{-n\Lambda^*(q)} E_{\tilde{\mu}} \left[e^{-\eta\sqrt{n\Lambda''(\eta)}\, W_n}\, 1_{W_n \geq 0} \right] = e^{-n\Lambda^*(q)} \int_0^\infty e^{-\eta\sqrt{n\Lambda''(\eta)}\, x} dF_n(x) \tag{2.10.246}$$

since $\hat{S}_n = q + \sqrt{\frac{\Lambda''(\eta)}{n}}\, W_n$. Let $\psi_n \triangleq \eta\sqrt{n\Lambda''(\eta)}$. By an integration by parts in (2.10.246) one obtains

$$J_n \mu_n(A) = \sqrt{2\pi} \int_0^\infty \psi_n^2 e^{-\psi_n x} [F_n(x) - F_n(0)]\, dx = \sqrt{2\pi} \int_0^\infty \psi_n e^{-t} \left[F_n\left(\frac{t}{\psi_n}\right) - F_n(0) \right] dt \tag{2.10.247}$$

Consider now

$$c_n \triangleq \sqrt{2\pi} \int_0^\infty \psi_n e^{-t} \left[\Phi\left(\frac{t}{\psi_n}\right) + \frac{m_3}{6\sqrt{n}} \left[1 - \left(\frac{t}{\psi_n}\right)^2 \right] \phi\left(\frac{t}{\psi_n}\right) - \Phi(0) - \frac{m_3}{6\sqrt{n}}\, \phi(0) \right] dt \tag{2.10.248}$$

Comparing (2.10.247) and (2.10.248), observe that the Berry-Esseen expansion (2.10.245) yields the relation $\lim_{n\to\infty} |J_n \mu_n(A) - c_n| = 0$. Moreoever, since

$$\sup_{x \geq 0} |\phi'(x)| < \infty, \quad \lim_{x \to 0} |\phi'(x)| = 0 \tag{2.10.249}$$

it follows by a Taylor expansion of $\Phi\left(\frac{t}{\psi_n}\right)$ and the dominated convergence theorem that

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} \sqrt{2\pi} \int_0^\infty \psi_n e^{-t} \left[\Phi\left(\frac{t}{\psi_n}\right) - \Phi(0) \right] dt =$$

$$= \lim_{n\to\infty} \sqrt{2\pi} \int_0^\infty e^{-t} \phi\left(\frac{t}{\psi_n}\right) dt = \sqrt{2\pi}\, \phi(0) \int_0^\infty e^{-t}\, dt = 1\,. \tag{2.10.250}$$

This completes the proof for the non-lattice case.

(b). In the lattice case (where the range of Y_i is $\left\{m\frac{d}{\sqrt{\Lambda''(\eta)}}\right\}_{m=-\infty}^{\infty}$) the Berry-Esseen expansion (2.10.245) is modified to

$$\lim_{n\to\infty}\left\{\sqrt{n}\sup_x\left|F_n(x)-\Phi(x)-\frac{m_3}{6\sqrt{n}}\,(1-x^2)\,\phi(x)-\phi(x)g\left(x,\frac{d}{\sqrt{\Lambda''(\eta)n}}\right)\right|\right\}=0 \qquad (2.10.251)$$

where $g(x,h)=\frac{h}{2}-(x \mod h)$ if $(x \mod h)\neq 0$ and $g(x,h)=-\frac{h}{2}$ if $(x \mod h)=0$ (see [12], page 513 or [22], page 171, Theorem 6). Thus, by adopting the argument above for the lattice case, one obtains

$$\lim_{n\to\infty} J_n\mu_n(A)=1+\lim_{n\to\infty}\sqrt{2\pi}\int_0^\infty \psi_n e^{-t}\left\{\phi\left(\frac{t}{\psi_n}\right)g\left(\frac{t}{\psi_n},\frac{\eta d}{\psi_n}\right)-\phi(0)g\left(0,\frac{\eta d}{\psi_n}\right)\right\}dt \qquad (2.10.252)$$

Since $\psi_n g\left(\frac{t}{\psi_n},\frac{\eta d}{\psi_n}\right)=g(t,\eta d)$ it follows that

$$\begin{aligned}\lim_{n\to\infty} J_n\mu_n(A) &= 1+\lim_{n\to\infty}\sqrt{2\pi}\int_0^\infty e^{-t}\left\{\phi\left(\frac{t}{\psi_n}\right)g(t,\eta d)-\phi(0)g(0,\eta d)\right\}dt\\ &= 1+\sqrt{2\pi}\,\phi(0)\int_0^\infty e^{-t}\left[g(t,\eta d)-g(0,\eta d)\right]dt. \qquad (2.10.253)\end{aligned}$$

The proof is completed by combining (2.10.253) with

$$\int_0^\infty e^{-t}[g(t,\eta d)-g(0,\eta d)]\,dt=\left\{\sum_{n=0}^\infty e^{-n\eta d}\right\}\int_0^{\eta d} e^{-t}(\eta d-t)\,dt=\frac{\eta d}{1-e^{-\eta d}}-1. \qquad (2.10.254)$$

□

Exercises:

2.10.1 (a). Let $A=[q,q+\frac{a}{n})$, where in the lattice case $\frac{a}{d}$ is restricted to be an integer. Prove that for any $a\in(0,\infty)$, both (2.10.243) and (2.10.244) hold with $J_n=\eta\sqrt{\Lambda''(\eta)\,2\pi n}\,e^{n\Lambda^*(q)}\frac{1}{1-e^{-\eta a}}$.
(b). As a consequence of part (a) above conclude that for any set $A=[q,q+b_n)$ both (2.10.243) and (2.10.244) hold for J_n as given in Theoren 2.10.2 as long as $\lim_{n\to\infty} nb_n=\infty$.

2.10.2 (a). Let $\eta>0$ denote the minimizer of $\Lambda(\lambda)$ and suppose that $\Lambda(\lambda)<\infty$ in some open interval around η. Based on exercise 2.10.1, deduce that the limiting distribution of S_n conditional upon $S_n\geq 0$ is Exponential(η) when X_1 has a non-lattice distribution.
(b). Suppose now that X_1 has a lattice distribution of span d and $1>\text{Prob}(X_1=0)>0$. Deduce now that the limiting distribution of $\frac{1}{d}S_n$ conditional upon $S_n\geq 0$ is Geometric(p) with $p=1-e^{-\eta d}$ (i.e., $\text{Prob}(S_n=kd|S_n\geq 0)\to pq^k$ for $k=0,1,2,\cdots$).

2.10.3 Consider a Neyman-Pearson test with constant threshold $\gamma \in (\overline{x}_0, \overline{x}_1)$ (see Section 2.7 for details). Suppose that $X_1 = \log \frac{d\mu_1}{d\mu_0}(Y_1)$ has a non-lattice distribution. Let $\lambda_\gamma \in (0,1)$ be the unique solution of $\Lambda_0'(\lambda) = \gamma$. Deduce from (2.10.243) that

$$\lim_{n\to\infty} \left\{ \alpha_n e^{n\Lambda^*_0(\gamma)} \lambda_\gamma \sqrt{\Lambda''(\lambda_\gamma)\, 2\pi n} \right\} = 1 \tag{2.10.255}$$

and

$$\lim_{n\to\infty} \left\{ \frac{e^{n\gamma}\alpha_n}{\beta_n} \right\} = \frac{1-\lambda_\gamma}{\lambda_\gamma} \,. \tag{2.10.256}$$

Chapter 3

Historical notes and references

Although much of the credit for the modern theory of large deviations and its various applications must go to Donsker and Varadhan, the topic is much older and references to the various aspects of it may be traced back to the early 1900's. Due to our own ignorance we will necessarily confine ourselves here to an incomplete list of references and historical credits. We hope to expand and correct this list at a later stage, and apologize to those who are not given due credit.

3.1 Chapter 2

The early development of large deviation bounds did not follow the order of our presentation. Statisticians, starting with Khinchin [18], have anaylsed various forms of Cramer's theorem for special random variables. See [27], [20] and [21] for additional references on this early work.

The first statement of Cramer's Theorem for distributions on R possessing densities is due to Cramer [7], who introduced the change of measure argument to this context. An extension to general distributions was done by Chernoff [6], who introduced the upper bound which was to carry his name. There exists a large body of literature concerning the applications of Cramer's theorem to the analysis of statistical tests, on which we keep silent.

Although Stirling's formula, which is at the heart of the combinatorial estimates of section 2.1, dates back at least to the 19-th century, the notion of types and bounds of the form of Lemmas 2.1.1-2.1.4 had to wait until information theorists discovered that they are useful tools for analyzing

the efficiency of codes. For early references we refer the reader to the excellent book of Gallager [12]. Our treatment of the source coding theorem in section 2.9 is a combination of the method of that book with the particular case treated by Bucklew in [4].

The credit for the extension of Cramer's theorem to the dependent case should definitely go to Gärtner [14] who considered the case in which $\mathcal{D}_\Lambda = \mathbb{R}^d$. Ellis [10] extended this result to the steep set up and the formulation of section 2.3 is nothing but an embellishment of his results.

The large deviations statements for Markov chains have a long history which is partially described in the historical notes of Chapter ??. The approach taken here is based in part on the ideas of Ellis [10].

The material in section 2.5 is a large deviations proof of the results in [22].

Gibb's conditioning principle has served as a driving force behind Ruelle and Landford's treatment of large deviations (without calling it by that name), [24], [25], [19]. The form of the Gibb's principle here was proved using large deviations methods (via the method of types) by Campenhout and Cover [5] and in greater generality by Csizar [8] and Stroock-Zeitouni [28].

No references yet for Stein's lemma.

The generalized maximum likelihood of section 2.8 was considered by Hoeffding [16], whose approach we basically follow-here. The extension to general state space presented in section ?? is due to Zeitouni and Gutman [30].

Finally, the refinements of the large deviations principles discussed in sections 2.10 and ?? follow [1], [21], although some of the methods are much older and may be found in Feller's book [11].

Bibliography

[1] R.R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *Ann. Math Statistics*, 38:1015–1027, 1960.

[2] T. Berger. *Rate Distortion Theory*. Prentice Hall, 1971.

[3] P. Billingsley. *Convergence of Probability Measures*. Wiley, 1968.

[4] J. A. Bucklew. *Large deviations techniques in decision, simulation, and estimation*. Springer, 1991.

[5] J.M. Van Campenhout and T.M. Cover. Maximum entropy and conditional probability. *IEEE Transc. Inf. Theory*, IT-27:483–489, 1981.

[6] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math., Statist.*, 23:493–507, 1952.

[7] H. Cramèr. Sur un nouveau théorème-limite de la théorie des probabilités. In *Actualités Scientifiques et Industrielles*, volume 3 of *Colloque consacré à la théorie des probabilités*, pages 5–23. Hermann, Paris, 1938.

[8] I. Csiszàr. I–divergence geometry of probability distributions and minimizations problems. *Ann. Probab.*, 3:146–158, 1975.

[9] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North Holland, Amsterdam, 1976.

[10] R. S. Ellis. Large deviations for a general class of random vectors. *Ann. Probab.*, 12:1–12, 1984.

[11] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley and Sons, New York, 1957.

[12] R.G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.

[13] F.G. Gantmacher. *The Theory of Matrices*. Chelsea, 1959.

[14] J. Gärtner. On large deviations from the invariant measure. *Theory Probab. Appl.*, 22:24–39, 1977.

[15] U. Grenander and G. Szego. *Toeplitz Forms and their Applications*. University of California Press, 1958.

[16] W. Hoeffding. On probabilities of large deviations. In *Proceedings of the Fifth Berkeley Symposium on mathematical Statistics and Probability*, pages 203–219. Univ. of California Press, 1965.

[17] I. Csizar, T.M. Cover and B.S. Choi. Conditional limit theorems under Markov conditioning. *IEEE Trans. Inf. theory*, IT-33:788–801, 1987.

[18] A.I. Khinchin. Uber einen neuen grenzwertsatz der wahrscheinlichkeitsrechnung. *Math Annalen*, 101:745–752, 1929.

[19] O.E. Landford. Entropy and equilibrium states in classical statistical mechanics. In A. Lenard, editor, *Statistical Mechanics and Mathematical problems*, volume 20 of *Lecture Notes in Physics*, pages 1–113. Springer, Berlin, 1973.

[20] Y. V. Linnik. On the probability of large deviations for the sums of independent variables. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 289–306, Berkeley, 1961. Univ. of California Press.

[21] V. V. Petrov. *Sums of independent random variables*. Springer, Berlin, 1975. Translated by A. A. Brown.

[22] R. Arratia, L. Gordon and M.S. Waterman. The Erdos-Renyi law in distribution for coin tossing and sequence matching. *Ann. of Statistics*, 1990.

[23] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[24] D. Ruelle. Correlation functionals. *J. math. Physics*, 6:201–220, 1965.

[25] D. Ruelle. A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule. *Comm. math. Phys.*, 5:324–329, 1967.

[26] E. Seneta. *Non Negative Matrices and Markov Chains.* Springer-Verlag, 1981.

[27] N. Smirnoff. Über warhscheinlichkeiten grosser abswichungen. *Rec. Sco. Math. Moscou*, 40:441–455, 1933.

[28] D.W. Stroock and O. Zeitouni. Micro canonical distributions, Gibbs' states, and the equivalence of ensembles. In R. Durett and H. Kesten, editors, *Festchrift in honour of F. Spitzer.* Birkhauser, 1990.

[29] T.M. Cover and J.B. Thomas. *Elements of Information Theory.* forthcoming, 1991.

[30] O. Zeitouni and M. Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inf. Theory*, 1991.